Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Structure-Based Phylogenomics Reveals the Evolutionary Reservoir of Plant Cysteine-Rich Peptides

Kenan Tan

tan@ipk-gatersleben.de

Leibniz Institute of Plant Genetics and Crop Plant Research https://orcid.org/0009-0006-4965-2773

Jiajie He

Leibniz Institute of Plant Genetics and Crop Plant Research

Twan Rutten

IPK Gatersleben

Amanda Camara

IPK Gatersleben https://orcid.org/0000-0002-3136-6633

Thorsten Schnurbusch

Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Corrensstr. 3 OT Gatersleben, D-06466 Seeland, Germany https://orcid.org/0000-0002-5267-0677

Article

Keywords: Cysteine-rich peptides, Structure-based phylogenomics, Evolutionary reservoir, Male fertility, Poaceae

Posted Date: August 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10486159/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Title:

Structure-Based Phylogenomics Reveals the Evolutionary Reservoir of Plant Cysteine-Rich Peptides

Kenan Tan[1*], Jiajie He[1], Twan Rutten[1], Amanda Souza Camara[1*], Thorsten Schnurbusch[1,2,§*]

[1]Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Corrensstraße 3, 06466 Seeland OT Gatersleben, Germany

[2]Faculty of Natural Sciences III, Institute of Agricultural and Nutritional Sciences, Martin Luther University Halle-Wittenberg, D-06120 Halle, Germany

[§]Cluster of Excellence on Plant Sciences (CEPLAS) 'SMART Plants for Tomorrow's Needs', Germany

[*]Correspondence: tan@ipk-gatersleben.de; camara@ipk-gatersleben.de; schnurbusch@ipk-gatersleben.de

ORCIDs

KT: 0009-0006-4965-2773

JH: 0009-0000-3325-0186

ASC: 0000-0002-3136-6633

TR: 0000-0001-5891-6503

TS: 0000-0002-5267-0677

Abstract

Cysteine-rich peptides (CRPs) are vital components in many prokaryotic and eukaryotic organisms. However, their remarkable sequence diversity has left them as a largely uncharted "dark proteome" within plant genomes. Here, we establish a structure-based phylogenomic approach to decode the evolutionary landscape of CRPs across 868 species. By integrating high-throughput structural clustering, we redefined the CRP superfamily into Major Superfamilies (MSs), Subfamilies (SFs), and Emerging Families (EFs), achieving a resolution that far surpasses traditional sequence-based methods. Our analysis reveals a massive expansion of CRPs in Poaceae, predominantly driven by the proliferation of EFs within transposable element-depleted, recombination-rich distal chromosomal regions. We propose that these EFs constitute an "evolutionary reservoir" characterized by structural and sequence plasticity. Crucially, we exemplified and functionally validated this model through the characterization of *Reduced Male Fertility 1* (*RMF1*), a Poaceae-specific CRP-lineage that evolved from the evolutionary reservoir into a fixed functional state. The transition of RMF1 is marked by a structural reorganization from a beta-sheet to a helical fold, accompanied by a shift from neutral drift to intense purifying selection. CRISPR/Cas9-mediated knockout of *RMF1* in rice and the identification of its TILLING mutants in wheat both lead to partly male sterility, confirming its essential role in cereal reproduction. Furthermore, as a proof-of-concept for breeding applications, we identified climate-adaptive CRP clusters on chromosome 6H and breeding-selected thionin-like clusters on 7H in European barley. Together, our study uncovers the evolutionary driving force underlying CRP expansion and provides a structure-based classification system which may facilitate harnessing the adaptive potential of the small-peptide dark proteome for future crop improvement.

Keywords: Cysteine-rich peptides; Structure-based phylogenomics; Evolutionary reservoir; Male fertility; Poaceae.

Introduction

Cysteine-rich peptides (CRPs) constitute a diverse class of small secreted signaling molecules that govern a vast array of biological processes in higher plants, encompassing antimicrobial defense, intercellular communication, and developmental regulation [1-3]. CRPs are frequently characterized by dense clusters in specific chromosomal regions, often exhibiting extensive copy number variations (CNVs) [2]. Based on their canonical functions, CRPs are generally categorized into three major functional groups: (i) biotic defense, including well-characterized subfamilies such as thionins [4-6], defensins [7-9]; (ii) developmental regulation, exemplified by the epidermal patterning factor/ epidermal patterning factor-like (EPF/EPFL) families [10] and lipid transfer proteins (LTPs) [11,12]; and (iii) reproductive signaling, such as the S-locus cysteine-rich

protein/S-locus protein 11 (SCR/SP11) [13,14], LURE [15], and rapid alkalinization factors (RALFs) [16] peptides. Despite this functional heterogeneity, CRPs share several defining structural characteristics: an N-terminal signal peptide for secretion, a compact protein scaffold, and a characteristic cysteine-rich domain containing approximately 4 to 16 conserved residues that facilitate disulfide bond formation [1,27]. However, our current understanding of CRPs is largely derived from the well-known families shown above, leaving the vast majority of the unknown CRPs systematically unexplored.

Due to this extreme primary sequence divergence, traditional genomic and evolutionary investigations have largely been confined to specific CRP subfamilies with well-defined roles, such as antimicrobial peptides [28], nodule-specific clusters [29], and canonical families including RALFs, LTPs, and thionins [30-32]. The field currently builds upon a foundational, comprehensive classification established nearly two decades ago, which first mapped the global landscape of the CRP superfamily[27]. The dramatic expansion of plant genomic resources and recent structural insights over the past twenty years present a timely opportunity to refine and update our understanding of CRP evolution across the plant kingdom. Conventional homology-based annotation approaches, such as BLAST searches [33] or HMM-based searches [34], are often challenged by the extreme sequence variability of CRPs, resulting in a very low frequency of well-annotated members in plant genomes [27]. Increasing evidence points to a pronounced decoupling between sequence and structure in these peptides: although their primary sequences evolve rapidly under diversifying selection, their 3D structural scaffolds remain highly conserved [35-37]. With the rapid advancement of computational methods and structure prediction algorithms, such as AlphaFold3 and ESMfold, protein structure prediction has become increasingly accurate and reliable [38-40], enabling the possibility of structure-based classifications. This structural conservation of CRPs represents an underexplored yet promising way for phylogenomic classification, with the potential to provide higher-resolution insights into CRP evolution than conventional sequence-based approaches.

The Poaceae is one of the most important plant families worldwide, including major crops that support global food security. Previous studies have suggested that Poaceae genomes harbor a large number of yet uncharacterized CRPs [27,28,41]. To date, only a few lineage-specific CRPs have been reported to be involved in crucial function such as *OsC6* [17], *OsRALF5* [42], while a large number of CRPs remain undiscovered with potentially important function for crop improvement. Beyond their classification and identification, however, the mechanisms driving the rapid expansion and diversification of CRP families in Poaceae remain incompletely understood. Poaceae genomes are characterized by their structural complexity and high transposable element (TE) content [43,44], yet it is still unclear whether CRP expansion is primarily associated with TE activity or driven by localized recombination hotspots in distal chromosomal regions. Resolving this question is critical for understanding the birth-and-death dynamics of gene families involved in plant–environment interactions, especially for rapidly evolving gene families such as CRPs superfamily.

In this study, we present a comprehensive structural phylogenomic analysis of the CRPs across 868 species, revealing the evolutionary landscape of these hypervariable peptides from a global perspective. By integrating high-throughput structural clustering with comparative genomics, we established a hierarchical classification system comprising Major Superfamilies (MSs), Subfamilies (SFs), and Emerging Families (EFs) (considered as "evolutionary reservoir"), effectively mapping the previously unannotated "dark proteome" of CRPs (Fig. 1E). Our genomic analyses reveal that CRP expansion in Poaceae is strongly associated with recombination-rich distal chromosomal regions, suggesting that these genomic environments facilitate the generation and maintenance of CRP diversity. Crucially, we validated our "evolutionary reservoir" hypothesis through the characterization of *Reduced Male Fertility 1* (*RMF1*), a Poaceae-specific lineage that evolved from the CRP reservoir into a fixed functional state. Utilizing CRISPR/Cas9-mediated gene editing and TILLING mutant, we demonstrate that loss of RMF1 leads to defective tapetum development in rice and partly male sterility in both rice and wheat, confirming its indispensable and conserved role during cereal reproduction. Furthermore, by leveraging a barley pangenome and GWAS, we identified a temperature-sensitive CRP cluster on chromosome 6H and a post-domestication Thionin-like cluster on 7H, highlighting the dual role of CRPs in natural adaptation and cereal reproduction. Overall, these results provide a transformative approach for decoding the functional diversity of CRPs and utilizing their adaptive potential for future breeding.

Materials and Methods

Data Source and Heuristic Identification of CRPs

Protein and coding sequences from 868 species, including bacteria, fungi, and plants, were retrieved from the OrthoDB v12.2 database [45]. To identify the maximum number of putative CRPs, we implemented a heuristic filtering pipeline based on previous studies[1,27], and the following structural and physicochemical constraints:

1. Cysteine content: between 2 and 30 residues;
2. Protein length: less than 300 amino acids;
3. Subcellular localization: presence of a N-terminal signal peptide and absence of any transmembrane domains (predicted via SignalP 6.0 [46] and deepTMHMM [47]).

To refine this dataset into HC CRPs, putative sequences were subjected to a profile-based search using HMMER v3.1b2 (hmmsearch with --cut_tc) against CRPs hmm files reported previously [27]. Orthologous groups (OGs) for HC CRPs were subsequently defined using OrthoFinder v2.5.5 [48].

Sequence Divergence and Diversity Analyses

Multiple sequence alignments of HC CRPs sequences were performed using MAFFT v7.490 [49]. To quantify the sequence-level constraints and conservation patterns, we calculated the cysteine residue count, nucleotide diversity (Pi), and positional entropy using custom scripts based on Biopython v1.86. These metrics were utilized to evaluate the degree of sequence conservation (0) versus hypervariability (1) within and across lineages. MCScanX [50] was used for identifying duplication events.

Structural Prediction and Fold-based Classification

Given the extreme sequence divergence of CRPs, a structure-based classification system was established. 3D atomic models for 8,224 representative HC CRPs were predicted in batch using a local installation of ESMFold [39]. Structural clustering and pairwise alignments were executed via FoldSeek [40] with the following optimized parameters:

- Clustering: -s 9.5, --alignment-type 1, --tmscore-threshold 0.5, --cluster-mode 1, -c 0.3, --cov-mode 1;
- Alignment: --alignment-type 1, -a 1, --tmscore-threshold 0.3.

Based on the resulting structural clusters, CRPs were categorized into a three-level evolutionary hierarchy: Major Superfamilies (MSs, $>10^3$ members, majorly cross-species), Sub Families (SFs, $>10^2$ members), and Emerging Families (EFs, remaining clusters, majorly lineage-specific) (Fig. S9). To determine the classification of all 58,513 HC CRPs, the remaining sequences without predicted structures were mapped back to these structural clusters via sequence alignment (--min-seq-id 0.3, -c 0.6, --cov-mode 1), enabling the assignment of each CRP to its corresponding structural family and yielding the total membership numbers for each hierarchical level. We retained low-pLDDT regions in the above-described analyses, considering that low-confidence regions may reflect functionally relevant intrinsic disorder and thus represent distinguishing structural features.

For higher structural prediction accuracy of our hypothesis-validating examples, CK, InterT and RMF1 were predicted by AlphaFold3 (AlphaFold Server, with default settings).

Conserved Domain Annotation and Phylogenetic Reconstruction

Functional domains were identified by scanning the NCBI Conserved Domain Database (CDD) [51] using BLAST+ v2.13.0 with an E-value threshold of $< 1 \times 10^{-5}$. For phylogenetic analysis, maximum-likelihood trees were reconstructed using IQ-TREE v2.2.2.6 [52], with the best-fit amino acid substitution model automatically selected by the built-in ModelFinder. The resulting trees were visualized and annotated to trace the expansion and contraction dynamics of CRP lineages.

Pangenome and Population Genetic Analysis in Barley

The evolutionary trajectories of CRPs within the barley lineage were investigated using a pangenome dataset of 76 barley accessions (comprising cultivars, landraces, and wild genotypes) [43]. Due to similarity of sequence interspecies, we use Orthofinder to cluster CRPs across 76 accessions. Selection Pressure: Pairwise non-synonymous to synonymous substitution rate ratios *Ka/Ks* were estimated using ParaAT2.0 [53]. Neutrality Tests: Tajima's D values were calculated using R package pegas [54] to identify signatures of selective sweeps or balancing selection across different populations.

Statistical Analysis

All statistical analyses were performed using R (v4.3.3). Pairwise comparisons between groups were conducted using two-tailed Student's t-tests or Wilcoxon rank-sum tests, as appropriate. To account for multiple hypothesis testing when making comparisons across multiple OGs, structural categories (MSs, SFs, EFs), and population groups, P-values were adjusted using the Benjamini-Hochberg false discovery rate (FDR) correction method via the `p.adjust()` function in R. For all analyses, an adjusted P-value < 0.05 was considered statistically significant, unless otherwise specified in Fig. legends. For specific pairwise comparisons with a priori hypotheses, unadjusted P-values are reported.

Structural Distance Calculation and Quantitative Structural Metrics

Pairwise structural similarity was quantified using TM-score (Foldseek pairwise alignment). A distance matrix D (D = 1 - TM-score) was constructed. 1280-dimensional feature vectors matrix were extracted using the pre-trained protein language model esm2_t33_650M_UR50D for subsequent PCA analysis. The dataset underwent quality control to verify metadata matching and remove inconsistent entries prior to downstream analysis.

We employed three distinct approaches to visualize the structural landscape based on the generated distance matrix and feature matrix:

- UMAP: Structural data were projected into a two-dimensional space using UMAP with the following parameters: n_neighbors = 30 and min_dist = 0.5.
- PCoA: Classical multidimensional scaling (via the cmdscale function) was applied to the distance matrix to generate linear embeddings.
- PCA: Principal Component Analysis was performed on feature matrix to obtain orthogonal linear projections.

To characterize relationships between protein families, the following metrics were calculated based on the distance matrix:

- Structural Centrality: The mean structural distance from each protein to all members within the EFs.

- Inter-family Divergence: The average pairwise distance between protein families.
- Within-family Compactness: The average pairwise distance among members within each individual family.

To interpret the structural landscape, we mapped the proteins onto the UMAP embedding based on their annotated functional domains from NCBI CDD (Conserved Domain Database). Statistical comparisons between protein families were conducted using the non-parametric Wilcoxon rank-sum test. All data processing and visualizations were performed in R, with Fig.s generated using ggplot2.

Disulfide Bond Analysis, Intrinsic Disorder and Binding Potential Prediction

To identify disulfide bonds, the spatial coordinates of cysteine sulfur atoms ($S\gamma$) were extracted from the predicted structures, and pairwise $S\gamma - S\gamma$ distances were calculated. Disulfide bonds were assigned when the sulfur–sulfur distance ranged from 1.8 Å to 2.5 Å, ensuring physically plausible connectivity while excluding potential modeling artifacts. Secondary structure elements surrounding validated cysteine residues were annotated using DSSP[55].

Sequence-based computational scanning was performed using AIUPred[56]. For each protein sequence, three functional metrics were calculated at a single-residue resolution: intrinsic disorder propensity, disordered binding potential, and linker propensity. A residue was classified as structurally disordered if its disorder propensity score exceeded the standard threshold of 0.5. Based on this residue-level classification, the global structural plasticity of each protein was quantified using the Fraction of Disordered Residues (FoDR), defined as follows:

$$FoDR = \frac{Number\ of\ Disordered\ Residues}{Total\ Sequence\ Length}$$

Concurrently, the potential molecular recognition features (MoRFs) mediating induced-folding interactions were identified via the binding potential score. To capture localized, high-affinity interaction "hotspots" that might otherwise be diluted by global averaging in short peptides, the Maximum Binding Score (Max_Binding) was extracted for each sequence.

Transcriptomic Integration and Genomic Context Analysis

To evaluate the functional relevance of CRPs expansion, we integrated public transcriptomic data covering 4 tissues (including anther, leaf, root, seed) and stress conditions (including heat, cold, salinity, and pathogen infection) from multiple BioProjects (e.g., PRJNA558196, PRJNA324116, PRJNA728113) [57]. Expression levels were normalized and clustered to identify tissue-specific or stress-responsive CRP families. Genome-wide synteny and collinearity analyses were performed using JCVI v1.4.4 [58]. The genomic distribution of

CRPs relative to TEs was analyzed by calculating the distance to the nearest TE and the TE density within 40-kb windows. The association between CRP clusters and recombination hotspots was assessed using published barley recombination maps [59]. Distribution of CRPs and TE was visualized by Circus v0.69.9 [60].

Temperature- sensitivity index (TSI) Calculation and GWAS Analysis

To evaluate the genetic basis of heat tolerance in barley, we used phenotypic data for three TSI traits from previous publication: grain number per spike (GNS), grain area (GA), and fertility rate (FR) across two contrasting environments (the extreme warm/dry year of 2018 and the normal year of 2020). By utilizing climate data from IPK record, a Best Linear Unbiased Prediction (BLUP) was calculated for each trait using a linear mixed model. Based on the BLUP values, a temperature sensitivity index was derived to represent the plasticity of each accession under thermal stress, which served as the input phenotype for subsequent association mapping.

Genome-wide association study (GWAS) was conducted to identify genomic loci associated with temperature sensitivity index. Genotype data are obtained from previous publication. High-quality SNPs from the barley pangenome (MAF > 0.05, missingness < 10%) were mapped against the cv. Morex V2 reference genome, with population structure and cryptic relatedness controlled using PCA and kinship matrix. The association analysis was performed using the MLMA model with software gcta v1.93.3 [61].

Plant Materials and Growth Conditions

Three independent CRISPR/Cas9-mediated knockout lines (BG103048A07, BG103408G04, and BG103409E09) in the rice (Oryza sativa L.) cv. Zhonghua 11 (ZH11) background were obtained from BIOGLE Co., Ltd. (Jiangsu, China). For wheat (F2 segregation population), two TILLING mutant lines (WCAD0457 and WCAD1723) of bread wheat (Triticum aestivum L.) cv. Cadenza were obtained from the Genetic Resources Unit (GRU) at the John Innes Centre (JIC, UK). All experimental trials were conducted under controlled environments at the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK, Gatersleben, Germany). Rice plants were cultivated in a Phytochamber under a diurnal temperature cycle of 28°C/24°C (day/night) with a 12-h photoperiod. Wheat plants were grown in a greenhouse maintained at 22°C/18°C (day/night) under a 16-h long-day photoperiod.

Fertility and Pollen Phenotyping

To evaluate reproductive success and yield components, the primary panicle (rice) or main spike (wheat) was harvested at physiological maturity. The total number of spikelets and filled grains per panicle/spike were recorded to calculate the fertility rate (the ratio of grains per spikelet). Male fertility was further assessed at the heading stage through two complementary methods. Pollen viability was determined via iodine-potassium iodide I_2-KI staining, where pollen grains were classified as fertile (deeply stained) or

sterile (unstained/malformed) under light microscopy. Additionally, total pollen quantity and physiological integrity were precisely quantified using an Ampha Z40 impedance flow cytometer (Amphasys AG, Switzerland) according to the manufacturer's established protocols. Statistical analyses were performed to determine significant differences in reproductive performance between the mutant lines and their respective wild-type controls.

Anther Fixation, Sectioning, and phenotyping

Anthers were collected and fixed in 4% formaldehyde in 50 mM phosphate buffer (pH 7.0) with 0.05% Triton X-100 for 72 h at 8°C. Following dehydration through a graded ethanol series (30%–100%) and a 10-min incubation in 100% propylene oxide, samples were gradually infiltrated and embedded in Spurr's resin to ensure proper orientation of the anther lobes, followed by polymerization at 60°C for 16 h. Serial semi-thin sections (1 μm thick) were prepared using a Reichert-Jung Ultracut S microtome, with consecutive section spacing dynamically adjusted (e.g., 2–10 μm) depending on the specific anther developmental stages. Sections were stained with 0.01% aqueous crystal violet, heat-fixed at 90°C, and imaged using a Zeiss Axio Scan Z1 slide scanner (10×/0.45 NA objective).

Results

Mapping the Pan-Plant Diversity and Expansion Patterns of the CRP Superfamily

Through an integrated heuristic and HMM-based identification pipeline (Fig. 1B, Supplementary Table S1, S2), we performed a kingdom-wide survey of CRPs spanning from prokaryotes and fungi to diverse green plant lineages. A total of 81,248 putative CRPs were identified, with 58,513 categorized as high-confidence (HC) peptide molecules after stringent hmmsearch. These HC CRPs were clustered into 500 distinct orthologous groups (OGs) using OrthoFinder, reflecting a vast and complex genetic library. An analysis of cysteine architecture revealed a discrete distribution of Cys residues, with the 8-Cys configuration emerging as the dominant architectural motif, followed by frequent clusters of 4, 6, 10, and 12-Cys patterns (Fig. 1A). This structural diversity at the primary sequence level suggests a versatile scaffold capable of supporting myriad biological functions.

To characterize the selective constraints acting on these hypervariable peptides, we calculated the amino acid diversity (Pi-aa) and positional entropy across the CRP bodies. We observed a significant spatial partitioning of peptide sequence conservation: The N-terminal signal peptides exhibited high levels of Pi and entropy, whereas the Cys-rich domains showed relatively lower values, indicating stronger purifying selection (Fig. 1A). However, the conservation within the Cys-rich domains was less pronounced than that observed in canonically conserved protein domains. Our results and previous reports [1,27] demonstrate that while the primary amino acid sequences between cysteine residues remain highly plastic, the number and

relative spacing of Cys residues are the primary evolutionary constraints. This "sequence-variable, pattern-conserved" signature highlights the limitations of conventional sequence-based alignment and necessitates a structural approach for accurate classification.

Phylogenomics of the CRP superfamily reveals a progressive evolutionary expansion spanning from ancestral prokaryotes to advanced angiosperms, with modern flowering plants displaying significantly higher complexity than early-diverging bryophytes and pteridophytes (Fig. 1C). Strikingly, within the monocot clade, the Poales lineage emerges as a singular evolutionary anomaly experiencing a massive, large-scale CRP proliferation. Deconstruction of this expansion profile (Fig. 1D) partitions the CRP proteomic space into five discrete OGs blocks. Two of these blocks remain deep-rooted and invariant across all green plants, underscoring their possible roles in core physiological housekeeping. Conversely, the remaining blocks manifest intense lineage-specific innovations, prominently bifurcating into independent expansions unique to Poales (predominantly Poaceae) and Brassicales (predominantly Brassicaceae). This singular, independent evolutionary burst in Poales strongly implies that CRP diversification potentially served as a critical adaptive engine for the lineage. Consequently, we targeted our subsequent investigations downstream on the Poaceae family to dissect the precise evolutionary mechanisms driving this rapid genetic expansion and classification.

A Structure-Based Hierarchical Classification Resolves the Evolutionary Complexity of Hypervariable CRPs

Given that CRP sequences are hypervariable and often elude accurate classification via primary sequence alignment, we developed a structure-based classification pipeline to redefine the CRPs superfamily. To control computational complexity, we first clustered all HC CRPs based on sequence identity (minimum sequence identity of 30% and minimum coverage of 60%), collapsing them into 8,224 representative sequences, which were subsequently subjected to 3D structure prediction using ESMFold. The quality of the predicted models was robust, with an average pLDDT score of approximately 70 (Fig. 2A, Supplementary Fig. S1C). The pLDDT scores of CPRs are comparatively lower than those of other well-characterized protein families, reflecting a current paucity of experimental research and the inherent structural complexity of this family. As expected, we observed a distinct spatial partitioning of model confidence: the Cys-rich domains consistently yielded pLDDT scores exceeding 75, while the flanking regions exhibited higher structural disorder (Fig. 2B). This structural heterogeneity aligns with our sequence diversity analysis, suggesting that the cysteine-rich domain maintains a rigid topological scaffold, likely reflecting the functional plasticity of CRPs.

By employing FoldSeek for easy-clustering and all-vs-all structural alignment, we established a three-tiered hierarchical classification system consisting of 15 Major Superfamilies (MSs), 57 Subfamilies (SFs), and

1,368 Emerging Families (EFs) (Supplementary Table S3, S4). EFs possess the lowest average pLDDT scores, indicating that they may adopt a more flexible or disordered structural profile compared to MSs and SFs (Fig. 2A, Supplementary Fig. S7A). To visualize the structural landscape, we performed UMAP dimensionality reduction based on a global matrix D (1 – TM-score). The resulting UMAP projection demonstrated a remarkable convergence of structural and functional categories; for instance, multiple Major Superfamilies (MS_C1, C2, C10, C11, and C14) associated with LTPs clustered within the same structural landscape despite their high sequence divergence (Fig. 2C, 2G, Supplementary Fig. S1A). Other groups, including GASA, HPS, and Stig1, also formed distinct, tightly-packed clusters, validating the sensitivity of our structure-based approach (Fig. 2C, 2G, Supplementary Fig. S1A). Intriguingly, many EFs exhibited a highly dispersed distribution, intermingling with established MSs and SFs, implying pronounced structural plasticity within these newly emerging lineages. To systematically decode the functional and structural evolutionary trajectories underlying this plasticity, we mapped the proteomic space of the CRP superfamily into a landscape, intersecting the Fraction of Disordered Residues (FoDR) with the predicted Maximum Binding Score (Fig. 2E). Density-based analysis revealed a striking evolutionary differentiation: MSs and SFs conformed to a bimodal distribution, partitioning into either highly rigid, structurally constrained scaffolds (low FoDR <0.5, low max binding <0.5) or evolutionarily optimized flexible protein (low FoDR, high max binding). Conversely, EFs exhibited a pronounced shift toward the upper-right quadrant, characterized by elevated FoDR and high binding potential (Fig. 2E). Furthermore, within the EFs characterized by low pLDDT values, several genes (*AtRALF6/7/16/36/37* [19]) reported to possess important functions have been identified (Fig. 6A, Supplementary Fig. S7A, S7B). Given that low-confidence regions may reflect functionally relevant intrinsically disordered regions (IDRs), all proteins were retained for further analysis without filtering based on pLDDT scores. Potentially, these EFs leverage highly dynamic IDRs to execute molecular function via induced-folding or conformational selection. This structural flexibility inspired us to hypothesize that these proteins may represent a highly plastic “evolutionary reservoir” optimized for rapid adaptive innovation. Functional annotation via the Conserved Domain Database (CDD) revealed that while most MSs and SFs members correspond to established functional categories (e.g., LTPs, Defensins), the vast majority of EFs remain unannotated, representing a vast, unexplored "dark proteome" of plant peptides (Supplementary Fig. S1D, S7B, S7C).

Remarkably, a comparison with conventional methods revealed that sequence-based phylogeny failed to resolve these functional groupings, often scattering structural homologs across disparate branches (Fig. 2D, 2G, Supplementary Fig. S3). This highlights the resolution of structural proteomics in identifying distant homologs where sequence signal has been weakened by peptide evolution. Furthermore, by projecting unannotated CRPs into this structural space, we were able to assign putative functions to previously unknown genes based on their proximity to characterized folds (Supplementary Fig. S1A, S1B, S2).

Applying this approach to the lineage-specific expansions in Poaceae and Brassicaceae, the UMAP projection revealed a common structural core shared by both families, together with lineage-specific small and dispersed groups unique to each family (Fig. 2F, Supplementary Fig. 1D, Supplementary Table S5), with 71% and 68% of their specific CRPs being currently unannotated (Supplementary Fig. S1D). These divergent structural clusters represent independent evolutionary directions, prompting a deeper investigation into the specific mechanisms driving CRP expansion within the Poaceae family.

Recombination-Driven Expansion in Distal Chromosomal "Hotspots" Propels CRP Diversification in Poaceae

To elucidate the expansion dynamics of CRPs within the Poaceae, we reconstructed a high-resolution species phylogeny and mapped CRP family sizes across representative grasses and ancestral outgroups (*A. trichopoda*, *M. acuminata*, and *A. comosus*). Our analysis revealed two major lineage-specific expansion events during Poaceae evolution: the first occurred at the base of the Poaceae (divergence of *Pharus latifolius*), and the second coincided with the differentiation of the Triticeae tribe (Fig. 3A). Notably, both bursts were predominantly driven by the expansion of unannotated EFs, while the library of functionally characterized CRPs remained relatively stable (Fig. 3A). Comparative analysis of duplication modes using MCScanX and Chi-square tests revealed a shift in expansion mechanisms (Fig. 3C): whereas dispersed duplication dominated in the outgroups, the first expansion in Poaceae was primarily driven by singleton and tandem duplications. In contrast, the second event in Triticeae was characterized by a significant enrichment of proximal and whole-genome duplication (WGD, representing the WGD/segmental category from MCScanX) events (Fig. 3B, F), suggesting a transition in the genomic mechanisms governing CRP diversity.

We further investigated the genomic context of CRP expansion by examining the chromosomal distribution of CRPs and TEs in four representative species (*P. latifolius*, *O. sativa*, *Z. mays*, and *H. vulgare*) (Fig. 3D). In rice and barley, CRPs showed a strong preference for distal chromosomal regions, which are known to exhibit elevated recombination rates [43,59]. Permutation tests (1,000 iterations) comparing CRPs with randomly sampled genes confirmed that CRPs are significantly enriched in TE-depleted, recombination-active regions in *P. latifolius*, rice, and barley (Supplementary Fig. S4). These results support a predominant role of recombination in shaping CNV of CRPs, rather than a primary association with TE-rich regions. Notably, maize represents an exception: CRP expansion in maize is more closely associated with TE distribution and increased self-collinearity (Fig. 3D), likely reflecting lineage-specific TE bursts and genome expansion in this species.

Finally, we assessed the molecular evolution and structural stability of these expanded CRPs by comparing Poaceae-specific families to common orthogroups (Fig. 3E). Poaceae-specific CRPs exhibited significantly

higher *Ka/Ks* ratios, consistent with rapid sequence evolution and potential neofunctionalization under positive selection. This accelerated sequence variation was accompanied by pronounced structural divergence: Poaceae-specific CRPs displayed significantly lower TM-scores and lDDT (local Distance Difference Test) values, alongside markedly higher RMSD (Root Mean Square Deviation), compared to the conserved common set. This structural plasticity in lineage-specific CRPs likely facilitates the exploration of new functional space. To connect these macro-evolutionary patterns with micro-evolutionary processes, we subsequently utilized the barley pangenome to investigate the population-level forces shaping CRP diversity.

Integrated Pangenomic and Transcriptomic Landscapes Reveal the Evolutionary Plasticity of Barley CRPs

To investigate the genetic diversity of CRPs within the barley lineage, we identified 67,848 HC CRPs across the 76 barley pangenome accessions (averaging ~892 CRPs per genome). These sequences were clustered into 855 orthogroups (OGs) using OrthoFinder, including 306 EFs, 336 MSs, and 213 SFs. Across the pangenome, EFs, MSs, and SFs comprised 25,367, 25,845, and 16,590 CRPs, respectively (Supplementary Table S6). Notably, seven of the ten largest OGs belonged to the EF category, whereas only two were MSs and one was an SF, indicating that expansion within the largest CRP families predominantly occurred in Poaceae-specific EF lineages. While MSs and SFs OGs were well-annotated with canonical domains (e.g., LTPs, Thionins), the majority of EFs remained functionally uncharacterized and unannotated (Fig. 4A), highlighting a vast potential of lineage-specific novelty.

To decode the functional relevance of these CRPs, we integrated large-scale transcriptomic data covering various tissues and stress responses (biotic and abiotic). Based on their expression breadth and specificity, CRPs were clustered into 10 distinct clusters (C1–C10) (Fig. 4C). Specifically, clusters C1, C7, and C8 (127 genes) exhibited development-associated patterns; C10 and C3 (95 genes) were specifically upregulated in response to biotic infection (e.g., *Fusarium*); and C2, C5, and C9 (214 genes) demonstrated sensitivity to abiotic stresses. An alluvial analysis (Fig. 4B) revealed a significant dichotomy in transcriptional activity: despite the lack of functional annotations for most EFs, a substantial majority of them (61.47%) were transcriptionally active. This active proportion is comparable to that of SFs (61.34%) and only slightly lower than that of MSs (73.07%), implying the potential biological relevance of these uncharacterized genes. Notably, the conditionally expressed subset of EFs exhibited highly specialized responses to both biotic and abiotic stresses. Regarding its highly flexible structure and relatively recent expansion analyzed in the previous part, this suggests that these EFs may serve as an “evolutionary reservoir”, primed for activation to confer adaptive plasticity under specific environmental pressures.

Finally, we examined the impact of domestication and geographical origin on CRP evolution by calculating the *Ka/Ks* ratio and Tajima's D across the 76 accessions. We observed a clear evolutionary dichotomy: some CRPs remained highly conserved (low *Ka/Ks*) to maintain essential biological processes, while others exhibited signatures of rapid evolution (high *Ka/Ks*). Notably, cultivars displayed significantly lower *Ka/Ks* ratios in MSs and SFs compared to landraces and wild barleys, suggesting intensified purifying selection on established functional families during modern breeding (Fig. 4D). In contrast, landraces exhibited the highest *Ka/Ks* levels in MSs and SFs, potentially reflecting rapid local adaptation or relaxed selection during early domestication. Geographically, accessions from West Asia (the center of origination) exhibited significantly lower Tajima's D values, particularly in MSs and SFs groups, indicating potential selective sweeps or recent population expansions associated with adaptive CRP alleles (Fig. 4E). These findings demonstrate that while the core CRPs are preserved by purifying natural selection, the shell and cloud CRPs, especially within EFs, evolved to promote barley's adaptation to diverse ecological niches.

A Genomic Burst of Thionin-like Genes on Chromosome 7H: A Potential Target for Cereal Crop Improvement

To explore the breeding relevance of CRPs, we mapped the global distribution of identified CRPs onto the cv. Morex reference genome. The CRPs exhibited a highly non-random distribution, predominantly forming dense clusters in distal chromosomal regions which characterized by high recombination rates (Fig. 5A) [59]. We identified four prominent "CRP hotspots" located at the distal regions of 5H, and 6H, 7H, alongside a massive expansion at both ends of 7H. Smaller but distinct clusters were also observed on 1H, 2H, and 3H. This spatial enrichment in high-recombination zones strongly suggests that these regions serve as evolutionary hotspots, facilitating CNVs and structural diversification of CRPs, which aligns with the high evolutionary turnover and expansion observed in our cross-species analyses.

To link these genomic clusters with climate adaptation, we conducted a Genome-Wide Association Study (GWAS) targeting temperature-sensitive traits. By integrating phenotypic data from extreme weather years (warm/dry 2018 vs. normal 2020), we calculated the TSI using Best Linear Unbiased Prediction (BLUP) for traits including grain number per spike (GNS), grain area (GA), and fertility rate (FR) (Supplementary Fig. S5). A highly robust and consistent association peak, anchored by the lead SNP s6_2920808 ($P = 8.07*10^{-6}$), was mapped to the distal region of chromosome 6H. Intriguingly, this significant GWAS signal precisely co-localized with a major CRP cluster (Fig. 5E). Linkage disequilibrium (LD) block analysis delineated a 0.4 Mb (2.8 Mb – 3.2 Mb) physical interval surrounding this peak, which encompasses a total of 25 high-confidence genes, including 7 tandemly arrayed CRPs. To further resolve the genetic architecture of this polymorphic region, we examined the structural variants between two distinct parental assemblies (Hap1 - Morex and Hap2 – AIZU_6) marked by red lines indicating the precise genomic boundaries of the clustered CRPs in Fig. 5E). Despite the structural complexity of this locus, which complicates current

experimental validation, the robust GWAS signal suggests that the CRPs within this 6H cluster may play a pivotal role in modulating barley's reproductive resilience to thermal stress.

Finally, we performed a comparative analysis of CNVs across the identified CRP hotspots to detect signatures of post-domestication selective breeding. While most clusters remained stable, the distal 7H cluster exhibited a significant expansion in cultivars compared to wild barley accessions, with a markedly higher CNV frequency observed in European populations relative to Asian clades (Fig. 5B). Detailed collinearity analysis of this ~50 Mb region revealed a conserved core of four OGs, with the expansion primarily driven by OG-BarleyCDS90_24543, which is annotated as a Thionin-like protein (Fig. 5C, D). Given the established role of thionins in antimicrobial defense, the selective proliferation of this specific OG in European landraces and cultivars suggests a signature of selective breeding. This implicates that European breeding programs have favored the expansion of this Thionin-like cluster to bolster biotic resistance, positioning it as an important potential target for crop improvement.

From Plasticity to Function: a Poaceae-specific EFs member RMF1 Exemplifies the "Evolutionary Reservoir" Hypothesis.

Structural landscape analysis via UMAP revealed a striking hierarchical organization: while EFs, MSs, and SFs clustered discretely, a substantial subset of EFs exhibited spatial proximity to the other groups (Fig. 6A). To ensure the robustness of our structural mapping, we additionally performed PCA and PCoA to complement the UMAP results. Both methods recapitulated the broad global dispersion of EFs across the structural space (Fig. 6B, 6C). Quantitative metrics of matrix D further showed that EFs possess significantly lower structural disparity relative to both MSs and SFs compared to the disparity between those latter groups (Fig. 6D). Collectively, the minimal global structural distances observed for EFs (Fig. 6E) position them as a pivotal bridging group, representing a pool of structural prototypes between diversified protein families. Disulfide analysis further revealed this evolutionary divergence: MSs exhibit the highest density of high-confidence disulfide bonds to lock their rigid functional folds. Conversely, EFs display a stark paucity of these stable linkages but the highest overall count of low-confidence, flexible disulfide pairs, reflecting an unconstrained conformational evolutionary playground free for exploration (Fig. 6F, G). Notably, two distinct tiers exist within the central EFs region: the basal EFs consist primarily of CPRs with unknown functions, whereas the upper EFs bear diverse annotations (predominantly RALF-like proteins); this suggests that, despite sharing structurally similar low-confidence features, the upper EFs may already possess distinct functions (Fig. 6A). This central spatial arrangement in the map (Fig. 6A, B, C) and the younger origination of EFs phylogeny (Supplementary Fig. S9), integrated with their predominantly flexible and disordered architecture, lead us to hypothesize that EFs function as an "evolutionary reservoir", a plastic structural pool from which novel CRPs arise and eventually stabilize into defined functional states. We propose that lineage-specific EFs constitute an "evolutionary reservoir" within the CRP superfamily. Unlike

established CRP families characterized by conserved folds and strong purifying selection, EFs exhibit relaxed structural constraints, elevated evolutionary rates, and extensive conformational diversity. Many EFs may represent evolutionarily short-lived or rapidly diverging lineages, but a subset may undergo structural stabilization through the acquisition of stabilizing features such as optimized disulfide connectivity and novel topological arrangements. Once coupled to beneficial biological functions, these structurally stabilized variants become subject to purifying selection and eventually emerge as lineage-specific CRP families (Fig. 1E).

To test the hypothesis, we traced the evolutionary path of CRPs sharing a rare C-5-C-6-C-3-C-x-C-3-C-3-C-3-C motif, selecting three representative stages: Control/CK (with G-rich motif) within the reservoir (basal EFs), Intermediate/InterT (without G-rich motif) at the transition zone (Upper EFs), and RMF1 (stabilized) in the upper more stabilized pool of the EFs (Fig. 6A, 6B, 6C, Supplementary Table S7). Phylogenetic continuity of them is underscored by their cross-species distribution and exclusive mutual BLAST matches; the most basal CK-like homolog identified in our dataset originates in the ancient *Ginkgo biloba* (evm.model.chr7.68), transitions into the more widespread InterT across monocots and eudicots, and ultimately diverges into RMF1 exclusively within the Poaceae family (Supplementary Table S8, Supplementary Fig. S8). Analysis of selection pressure further corroborated this model: RMF1 exhibited the significantly lowest *Ka/Ks* ratio (close to 0) compared to InterT and CK (Fig. 6H), indicating that as a CRP departs from the structural reservoir and achieves potential functional fixation, it transitions from a state of high plasticity to one of intense purifying selection. We hypothesize that the migration of RMF1 out of the structural reservoir was driven by a fundamental topological shift. Structural modeling revealed that while reservoir-resident members (CK and InterT) maintain a prototypical "arrow-like" beta-sheet structure, RMF1 has evolved a distinct helical fold (Fig. 6J, Supplementary Fig. S10). Furthermore, we evaluated the potential for disulfide bond formation, a critical factor in structural stabilization. Our analysis revealed three predicted disulfide bridges in RMF1, compared to a single potential bond in InterT and none in CK. This marked disparity in cross-linking patterns suggests that disulfide-mediated structural constraint plays a pivotal role in the functional divergence and evolutionary adaptation of these protein families. This structural metamorphosis provides a mechanistic explanation for its distinct positioning on the UMAP landscape and its Poaceae-specific conservation. Such "sequence-structure decoupling" allows the CRP superfamily to maintain a stable cysteine scaffold in the reservoir while simultaneously exploring novel structure that facilitate specific biological interactions, ultimately leading to the birth of lineage-specific functional groups.

To verify the biological indispensability of these stabilized EFs, we obtained CRISPR/Cas9 knockout lines for *RMF1* in rice (*Oryza sativa* cv. ZH11). Phenotypic assays in 2024 and 2025 revealed that loss of *RMF1* leads to a reduction in fertility (Fig. 6K, L). Cytological characterization of the anthers demonstrated that

the tapetum in *rmf1* mutants fails to develop normally at Stage 10, resulting in a dramatic decrease in microspore/pollen density. By Stage 12, the remaining pollen grains exhibited aborted development and near-total loss of viability (Fig. 6K). To further examine whether the functional role of RMF1 is conserved across the Poaceae, we evaluated the phenotype of wheat (*Triticum aestivum* cv. Cadenza) TILLING mutants. Consistent with our observations in rice, the wheat mutants exhibited a significant reduction in grain setting and spike fertility (Fig. 6M, N). This cross-species phenotypic consistency, spanning both Oryza and Triticum lineages, underscores the fundamental importance of RMF1 in the reproductive development of cereal crops. These results provide compelling evidence that EFs are not merely 'genetic noise' but rather functional precursors that, upon functional fixation, become vital regulators of core processes in plants. Furthermore, the results also highlight that a vast cohort of lineage-specific, unannotated, and evolutionarily young CRPs is far from useless, instead harboring substantial, untapped evolutionary and biological potential.

Discussion

Our study provides a comprehensive structural phylogenomic map of the CRPs superfamily across 868 green plant species, establishing a promising strategy to uncover the evolutionary principles underlying the diversification of hypervariable peptides. By moving beyond primary sequence similarity toward 3D structural relationships, we identify a Poaceae-specific burst in CRP expansion largely associated with the emergence of novel EFs with mostly unknown functions. We propose that recombination-rich distal genomic regions provide favorable evolutionary environments that facilitate CRP diversification by promoting sequence innovation and copy-number variation. Crucially, functional validation of the RMF1 lineage provides empirical support for the evolutionary reservoir hypothesis. Our findings demonstrate that CRP lineages emerge from structurally diverse reservoirs and can undergo functional stabilization and evolve into conserved regulators of essential processes, as exemplified by RMF1-mediated cereal reproduction. The identification of stabilized functional groups, together with climate-associated and breeding-selected CRP loci on chromosomes 6H and 7H, highlights CRPs as a treasure trove for adaptive and agronomic innovation. Collectively, our results reveal how the interplay between structural plasticity and recombination-associated genome dynamics enables cereals to maintain a flexible reservoir of peptide diversity with potential roles in environmental adaptation and evolutionary innovation.

Functionally associated with reproductive signaling and antimicrobial defense [1], CRPs exhibit accelerated evolutionary rates and pronounced sequence divergence that far exceed those of canonical gene families. Based on our results and previous research, we propose that CRPs undergo a unique "sequence-structure decoupling" evolutionary trajectory: While conserved disulfide bonds anchor their core architectural

stability, hypervariable sequences leverage structural plasticity to drive neofunctionalization. Consequently, their evolutionary innovation depends less on sequence conservation than on conformational dynamics; the apparent structural "noise" of highly disordered, rapidly mutating regions serves as an active evolutionary playground for exploring novel molecular interactions. This extreme plasticity makes conventional homology-based methods, such as HMM or BLAST, quite ineffective for comprehensive classification. However, the integration of newly released structure prediction with fold-based clustering has ushered in the era of structural phylogenomics [38-40]. By prioritizing 3D fold conservation over primary sequence identity, this approach provides superior resolution for resolving the complex evolutionary history of CRPs and offers a robust, scalable strategy for investigating other unknown proteomes or hypervariable gene families across the tree of life.

The central positioning and pronounced dispersion of EFs within the global PCA/PCoA space, together with their discrete clustering in UMAP, likely reflect their intrinsically disordered nature, placing them at comparable structural distances from both MSs and SFs. This geometric pattern is consistent with multiple independent lines of evidence, including phylogenomic, evolutionary, and functional analyses, collectively supporting the hypothesis that EFs constitute a highly dynamic "evolutionary reservoir" within the CRP superfamily. This reservoir is characterized by flexible structures, recent lineage-specific emergence, relaxed selective constraints, and elevated variation rates, thereby serving as a hotspot for structural innovation. Within this conceptual idea, MSs and SFs may represent ancestral, structurally constrained templates, whereas EFs may have originated from them via CNVs and subsequent rapid structural diversification driven by functional redundancy. In contrast to the classical Ohno model [62], the high structural and functional plasticity of CRPs suggests that emerging EFs are not merely redundant duplicates but can explore diverse functional landscapes, reflecting a form of sequence–structure decoupling that may act as an evolutionary bet-hedging strategy. Empirical evidence from the RMF1 lineage supports this reservoir-to-function model, where transition from the amorphous reservoir state to a fixed functional gene is accompanied by structural reorganization from a flexible β-sheet–rich scaffold to a helix-stabilized fold reinforced by three disulfide bonds, together with a shift from near-neutral evolution to strong purifying selection ($Ka/Ks < 1$). The reduced fertility observed in rice and wheat *rmf1* mutants further indicates that once EFs become genetically fixed and functionally integrated, they may evolve into indispensable regulators of growth and development. Nevertheless, it remains unclear which selective pressures drive EF stabilization, whether the reservoir-to-function transition is reversible, and to what extent this model is generalizable across lineages, warranting further comparative studies of additional EFs.

The unique distribution of RALF members within the "evolutionary reservoir" unveils a critical conceptual leap (Supplementary Fig. S7A, B). Unlike rigidly folded CRPs like LTP, thionins or defensins, RALF peptides operate as highly flexible IDRs whose biological execution fundamentally requires conformational

plasticity rather than a static single-chain crystal structure. Landmark structural analyses have established that RALF peptides act as multipronged signaling hubs; their bioactive, disulfide-stabilized folds are strictly contextualized and templated upon encountering diverse homotypic or heterotypic receptor complexes, such as cell-wall LRX proteins[63] or membrane-bound FER-LLG1 assemblies[64]. While structure prediction algorithms that model proteins in isolation typically assign low pLDDT scores to this ligand-free, flexible state, this conformational freedom may instead enable RALFs to engage distinct receptors through mutually exclusive binding modes across different pathways. This strongly suggests that evolutionarily young CRP families (e.g. EFs) remain functionally primed and possess potential neofunctionalization capacity, despite their lack of stable 3D folds (low pLDDT) and high structural disorder. The distinct spatial location of RALF and RMF1 illustrate two divergent but equally successful neofunctionalization trajectories: a dynamic signaling route that retains a flexible, low-pLDDT disordered architecture to manage complex communication (RALF), and a structural stabilization route where macroscopic tertiary overhauls lock the peptide into a rigidly constrained, autonomous fold under intense purifying selection to secure absolute function (RMF1). Thus, these findings indicate that CRPs within the "evolutionary reservoir" may not be merely transient or dispensable components, even when characterized by structural disorder and obscure selection signatures. The continuous existence of certain lineages (e.g., CK) from ancestral to novel species provides evidence of their underlying biological significance, despite their high sequence variability. Investigating these unknown CRPs holds substantial potential for advancing our understanding of peptide-mediated intercellular and inter-organismal communication, as well as the underlying principles of dynamic genome evolution.

Despite the massive amplification of TEs characteristic of Poaceae genomes, the expansion of the CRPs superfamily appears to be largely independent of TE-mediated transposition. By localizing to TE-depleted gene-rich regions (euchromatin)—particularly within the distal regions of chromosomes—CRPs clusters may bypass the repressive effects of epigenetic silencing (e.g., DNA methylation) typically associated with TE-rich heterochromatin. This spatial positioning may preserve the transcriptional activation essential for rapid adaptive responses. Furthermore, the elevated rates of unequal crossing-over in these distal recombination hotspots provide a mechanism for CNVs generation, which may explain the dramatic expansion bursts of CRPs throughout the evolutionary history of green plants, and most notably, their lineage-specific expansion within the Poaceae. From the perspective of germplasm utilization, the lineage-specific expansion of the thionin-like cluster at the distal end of chromosome 7H in European barley cultivars may reflect the influence of geographical selective breeding. This divergent expansion could be related to adaptation against distinct pathogenic landscapes characteristic of European agro-ecosystems. However, the internal architecture of these CRPs hotspots—defined by high CNVs and dense clustering—presents significant challenges for traditional functional verification. Given the potential for functional

redundancy within these clusters, conventional single-gene knockouts may fail to produce observable phenotypes. Future CRP research may benefit from transitioning toward cluster-level approaches, leveraging long-read sequencing to resolve complex genomic repeats and multiplexed CRISPR-based deletions to dissect entire functional clusters. Such integrated approaches will be important for decoding the uncharacterized CRPs (e.g., lineage-specific EFs) within the barley pangenome and for potentially harnessing these peptides for engineering climate-resilient and disease-resistant crops.

Data Availability

Transcriptome datasets were obtained from publicly available resources, including anther (PRJNA558196), root (PRJEB50400), leaf (PRJNA744021), seed (PRJNA975859), dry tissues (PRJEB12540, PRJEB40905), salt stress (PRJNA489775, PRJNA781996), cold stress (PRJNA767196), heat stress (PRJNA324116), infection (PRJNA728113), and osmotic stress (PRJNA679445) [57]. The 76 barley pan-genome sequences and corresponding gene annotations were retrieved from a previously published dataset [43]. Ortholog sequence was obtained from OrthoDB (v12) [45].

Author contributions:

KT: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Visualization, Writing – Original Draft.

JH, TR, MM: Data Collection (Plant Phenotyping).

ASC: Conceptual Guidance, Writing - Review & Editing.

TS: Funding Acquisition, Supervision, Project Administration, Writing – Review & Editing.

Acknowledgments:

We thank publicly available databases for providing valuable resources that supported this study. We also thank Roop Kamal and Yongyu Huang for assistance with phenotypic data collection. This work was supported by the China Scholarship Council (CSC) to K.T., and by the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) through institutional core funding and infrastructure support.

Competing interests

The authors declare no competing interests.

Reference

1 Marshall, E., Costa, L. M. & Gutierrez-Marcos, J. Cysteine-rich peptides (CRPs) mediate diverse aspects of cell–cell communication in plant reproduction and development. *Journal of experimental botany* **62**, 1677-1686 (2011).

2 Silverstein, K. A., Graham, M. A., Paape, T. D. & VandenBosch, K. A. Genome organization of more than 300 defensin-like genes in Arabidopsis. *Plant physiology* **138**, 600-610 (2005).

3 Takeuchi, H. The role of diverse LURE-type cysteine-rich peptides as signaling molecules in plant reproduction. *Peptides* **142**, 170572 (2021).

4 Bohlmann, H. & Apel, K. Thionins. *Annual review of plant biology* **42**, 227-240 (1991).

5 Bohlmann, H. & Broekaert, W. The role of thionins in plant protection. *Critical reviews in plant sciences* **13**, 1-16 (1994).

6 Saberi Riseh, R., Fathi, F., Vatankhah, M. & Kennedy, J. F. Thionins: potential use in plant defense against pathogens. *Plant Molecular Biology* **115**, 77 (2025).

7 Lacerda, A. F., Vasconcelos, É. A., Pelegrini, P. B. & Grossi de Sa, M. F. Antifungal defensins and their role in plant defense. *Frontiers in microbiology* **5**, 116 (2014).

8 Parisi, K. *et al.* in *Seminars in cell & developmental biology.* 107-118 (Elsevier).

9 Thomma, B. P., Cammue, B. P. & Thevissen, K. Plant defensins. *Planta* **216**, 193-202 (2002).

10 Hara, K., Kajita, R., Torii, K. U., Bergmann, D. C. & Kakimoto, T. The secretory peptide gene EPF1 enforces the stomatal one-cell-spacing rule. *Genes & development* **21**, 1720-1725 (2007).

11 Kader, J.-C. Lipid-transfer proteins in plants. *Annual review of plant biology* **47**, 627-654 (1996).

12 Lev, S. Non-vesicular lipid transport by lipid-transfer proteins and beyond. *Nature reviews Molecular cell biology* **11**, 739-750 (2010).

13 Schopfer, C. R., Nasrallah, M. E. & Nasrallah, J. B. The male determinant of self-incompatibility in Brassica. *Science* **286**, 1697-1700 (1999).

14 Suzuki, G. *et al.* Genomic organization of the S locus: identification and characterization of genes in SLG/SRK region of S9 haplotype of Brassica campestris (syn. rapa). *Genetics* **153**, 391-400 (1999).

15 Zhong, S. *et al.* Cysteine-rich peptides promote interspecific genetic isolation in Arabidopsis. *Science* **364**, eaau9564 (2019).

16 Sharma, A. *et al.* Comprehensive analysis of plant rapid alkalization factor (RALF) genes. *Plant physiology and biochemistry* **106**, 82-90 (2016).

17 Zhang, D. *et al.* OsC6, encoding a lipid transfer protein, is required for postmeiotic anther development in rice. *Plant physiology* **154**, 149-162 (2010).

18 Li, X. *et al.* Overexpression of OsDT11, which encodes a novel cysteine-rich peptide, enhances drought tolerance and increases ABA concentration in rice. *Plant molecular biology* **93**, 21-34 (2017).

19 Zhong, S. *et al.* RALF peptide signaling controls the polytubey block in Arabidopsis. *Science* **375**, 290-296 (2022).

20 Takeuchi, H. & Higashiyama, T. Tip-localized receptors control pollen tube growth and LURE sensing in Arabidopsis. *Nature* **531**, 245-248 (2016).

21 Hsiao, P.-Y., Cheng, C.-P., Koh, K. W. & Chan, M.-T. The Arabidopsis defensin gene, AtPDF1. 1, mediates defence against Pectobacterium carotovorum subsp. carotovorum via an iron-withholding defence system. *Scientific Reports* **7**, 9175 (2017).

22 Gu, T. *et al.* An endophytic fungus interacts with the defensin-like protein OsCAL1 to regulate cadmium allocation in rice. *Molecular Plant* **17**, 312-324 (2024).

23 Luo, J.-S. *et al.* A defensin-like protein drives cadmium efflux and allocation in rice. *Nature communications* **9**, 645 (2018).

24 Hara, K. *et al.* Epidermal cell density is autoregulated via a secretory peptide, EPIDERMAL PATTERNING FACTOR 2 in Arabidopsis leaves. *Plant and Cell Physiology* **50**, 1019-1031 (2009).

25 Abrash, E. B. & Bergmann, D. C. Regional specification of stomatal production by the putative ligand CHALLAH. *Development* **137**, 447-455 (2010).

26 Abrash, E. B., Davies, K. A. & Bergmann, D. C. Generation of signaling specificity in Arabidopsis by spatially restricted buffering of ligand–receptor interactions. *The Plant Cell* **23**, 2864-2879 (2011).

27 Silverstein, K. A. *et al.* Small cysteine‐rich peptides resembling antimicrobial peptides have been under‐predicted in plants. *The Plant Journal* **51**, 262-280 (2007).

28 Ma, H. *et al.* Evolution of antimicrobial cysteine-rich peptides in plants. *Plant Cell Reports* **42**, 1517-1527 (2023).

29 Zorin, E. A. *et al.* A variable gene family encoding nodule-specific cysteine-rich peptides in pea (Pisum sativum L.). *Frontiers in plant science* **13**, 884726 (2022).

30 Lu, M. *et al.* Genome-wide identification of the RALF gene family in Camellia oleifera Abel. and the potentiality of CoRALF50 in pollen tube growth. *Plant Science* **359**, 112603 (2025).

31 Lu, Y. *et al.* Pan-genomic insights into LTP gene family evolution across diploid cotton species. *Frontiers in Plant Science* **16**, 1691339 (2025).

32 Yan, M. *et al.* Genome-wide identification and expression analysis of thionin family in rice (Oryza sativa) and functional characterization of osTHION15 in drought stress and ABA stress. *International Journal of Molecular Sciences* **26**, 3447 (2025).

33 Camacho, C. *et al.* BLAST+: architecture and applications. *BMC bioinformatics* **10**, 1-9 (2009).

34 Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic acids research* **39**, W29-W37 (2011).

35 Shafee, T. M., Lay, F. T., Phan, T. K., Anderson, M. A. & Hulett, M. D. Convergent evolution of defensin sequence, structure and function. *Cellular and Molecular Life Sciences* **74**, 663-682 (2017).

36 Bauer, F. *et al.* Structure determination of human and murine β‐defensins reveals structural conservation in the absence of significant sequence similarity. *Protein Science* **10**, 2470-2479 (2001).

37 Van der Weerden, N. L. & Anderson, M. A. Plant defensins: common fold, multiple functions. *Fungal Biology Reviews* **26**, 121-131 (2013).

38 Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493-500 (2024).

39 Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123-1130 (2023).

40 Van Kempen, M. *et al.* Fast and accurate protein structure search with Foldseek. *Nature biotechnology* **42**, 243-246 (2024).

41 Wu, J. *et al.* Evolution of the defensin-like gene family in grass genomes. *Journal of genetics* **95**, 53-62 (2016).

42 Kim, H.-J. *et al.* Pollen-Specific OsRALF5 Regulates Pollen Tube Growth in Oryza sativa. *Journal of Plant Biology* **69**, 35-47 (2026).

43 Jayakodi, M. *et al.* Structural variation in the pangenome of wild and domesticated barley. *Nature*, 1-9 (2024).

44 Schnable, P. S. *et al.* The B73 maize genome: complexity, diversity, and dynamics. *science* **326**, 1112-1115 (2009).

45 Kuznetsov, D. *et al.* OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. *Nucleic acids research* **51**, D445-D451 (2023).

46 Teufel, F. *et al.* SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nature biotechnology* **40**, 1023-1025 (2022).

47 Hallgren, J. *et al.* DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks. *biorxiv*, 2022.2004. 2008.487609 (2022).

48 Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome biology* **20**, 238 (2019).
49 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution* **30**, 772-780 (2013).
50 Wang, Y. *et al.* MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic acids research* **40**, e49-e49 (2012).
51 Wang, J. *et al.* The conserved domain database in 2023. *Nucleic acids research* **51**, D384-D388 (2023).
52 Minh, B. Q. *et al.* IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Molecular biology and evolution* **37**, 1530-1534 (2020).
53 Zhang, Z. *et al.* ParaAT: a parallel tool for constructing multiple protein-coding DNA alignments. *Biochemical and biophysical research communications* **419**, 779-781 (2012).
54 Paradis, E. pegas: an R package for population genetics with an integrated–modular approach. *Bioinformatics* **26**, 419-420 (2010).
55 Hekkelman, M. L., Salmoral, D. Á., Perrakis, A. & Joosten, R. P. DSSP 4: FAIR annotation of protein secondary structure. *Protein Science* **34**, e70208 (2025).
56 Erdős, G. & Dosztányi, Z. AIUPred: combining energy estimation with deep learning for the enhanced prediction of protein disorder. *Nucleic Acids Research* **52**, W176-W181 (2024).
57 Zhao, J. *et al.* BarleyOmics: A comprehensive multi-omics database of barley. *Molecular Plant* **18**, 1245-1248 (2025).
58 Tang, H. *et al.* JCVI: A versatile toolkit for comparative genomics analysis. *Imeta* **3**, e211 (2024).
59 Dreissig, S., Mascher, M. & Heckmann, S. Variation in recombination rate is shaped by domestication and environmental conditions in barley. *Molecular biology and evolution* **36**, 2029-2039 (2019).
60 Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome research* **19**, 1639-1645 (2009).
61 Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *The American journal of human genetics* **88**, 76-82 (2011).
62 Innan, H. & Kondrashov, F. The evolution of gene duplications: classifying and distinguishing between models. *Nature Reviews Genetics* **11**, 97-108 (2010). https://doi.org/10.1038/nrg2689
63 Moussu, S. *et al.* Structural basis for recognition of RALF peptides by LRX proteins during pollen tube growth. *Proceedings of the National Academy of Sciences* **117**, 7494-7503 (2020).
64 Xiao, Y. *et al.* Mechanisms of RALF peptide perception by a heterotypic receptor complex. *Nature* **572**, 270-274 (2019).

Figure

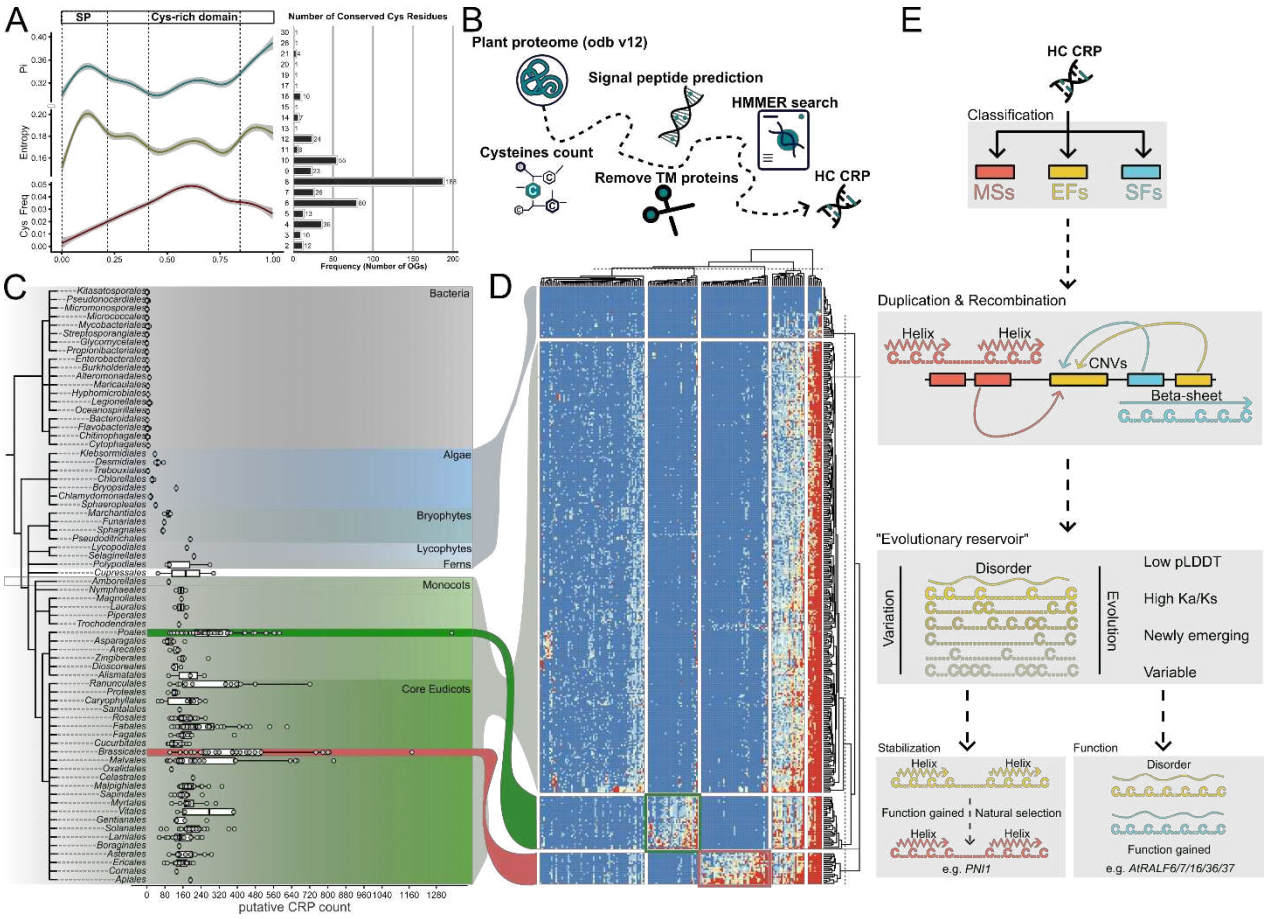


Fig. 1 Systematic identification and evolutionary expansion of CRPs across Viridiplantae. (A) The left figure shows structural features and conservation patterns of the CRP superfamily. The upper panel shows canonical CRP gene architecture. The lower panels depict cysteine residue distribution and conservation across HC CRPs, quantified by amino acid diversity (Pi-aa) and Shannon entropy, and cysteine positional frequency. The right figure shows frequency of OGs with different Cys residues number. (B) Overview of the computational approach for identifying putative and high-confidence (HC) CRPs, integrating sequence-based searches with structural filtering across 868 plant species. (C) Taxonomic distribution of identified CRPs across major plant orders, highlighting their extensive expansion within angiosperms. (D) Comparative distribution of CRP orthogroups across plant lineages. The heatmap illustrates lineage-specific expansion patterns, with Poaceae- and Brassicaceae-enriched orthogroups highlighted. (E) A proposed evolutionary model of the "evolutionary reservoir". (Top) Gene duplication and unequal recombination driven by copy number variations (CNVs) seed the structural components (e.g., helices and beta-sheets). (Middle) The EFs act as an "evolutionary reservoir," characterized by structural disorder, low pLDDT scores, high Ka/Ks ratios, and rapid, lineage-specific emerging variations. (Bottom left) Stabilization pathway: Under natural selection, specific disordered sequences undergo structural stabilization, rigidifying into defined secondary structures (e.g., helices) and gaining novel functions, as exemplified by RMF1. (Bottom left) Functional retention pathway: Alternatively, some lineages retain their intrinsically disordered conformations while successfully acquiring distinct biological functionalities, such as the *AtRALF6/7/16/36/37* clusters.

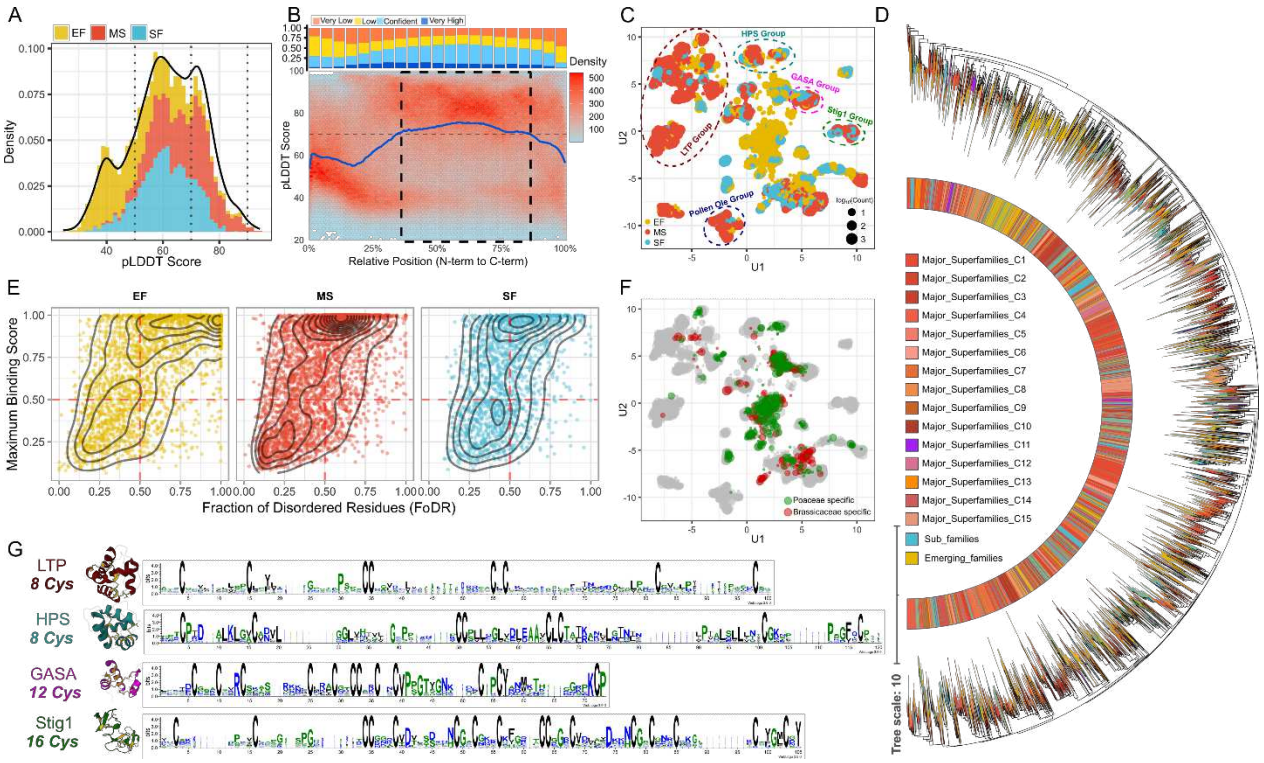


Fig. 2 Structural phylogenomics of high-confidence (HC) CRPs across the plant kingdom. (A) Distribution of predicted structure confidence (pLDDT scores) for HC CRPs across MSs, SFs and EFs. (B) Positional distribution of pLDDT scores along CRP sequences, with cysteine-rich region highlighted. (C) UMAP projection of HC CRPs based on structural features, colored by EFs, MSs, and SFs classifications. (D) Sequence-based phylogenetic tree of all HC CRPs. (E) Distribution landscape of EFs, MSs and SFs across conformational and interaction spaces, with kernel density estimation and red dashed lines as threshold. (F) UMAP projection highlighting lineage-specific CRPs, with Brassicaceae- and Poaceae-enriched clusters indicated. (G) Representative structures of four well-characterized CRP families, along with their cysteine-rich domain sequences.

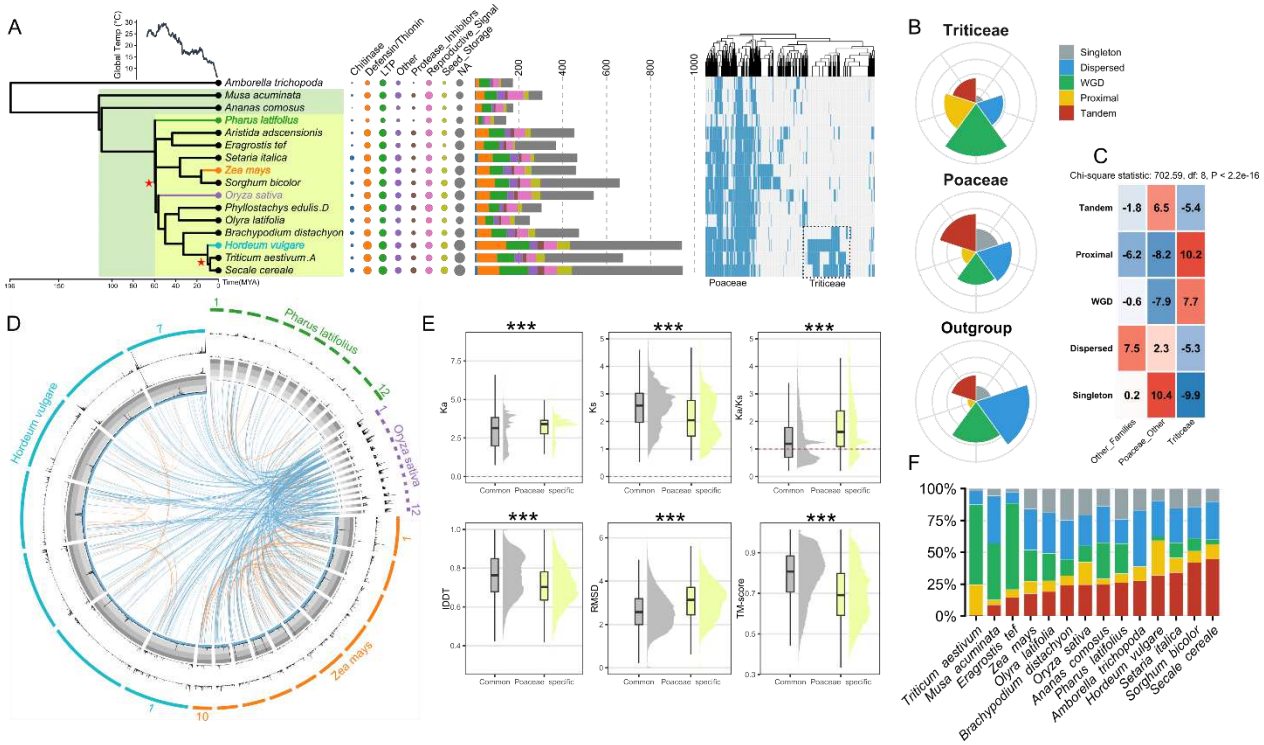


Fig. 3 Expansion patterns of HC CRPs in the Poaceae family. (A) Overview of CRP expansion across representative Poaceae species. From left to right: species phylogeny, proportions of annotated HC CRPs, total HC CRP counts per species, and OGs distribution heatmap. (B) Proportions of different duplication modes in Triticeae, Poaceae, and outgroup species. (C) Chi-square test of duplication mode distributions among Triticeae, Poaceae, and outgroup species. (D) Circos plot showing collinearity among four representative species. From outer to inner tracks: chromosomes, distribution of putative CRPs, distribution of HC CRPs, and distribution of TEs. Blue lines indicate interspecies collinearity, and yellow lines indicate intraspecies duplication events. (E) Comparative analysis of evolutionary and structural metrics, including Ka, Ks, Ka/Ks, IDDT, RMSD, and TM-score. Statistical significance was assessed using Student's t-test. (F) Distribution of duplication modes across different species. P-values were adjusted for 6 independent tests using the Benjamini-Hochberg method. FDR-adjusted $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$.

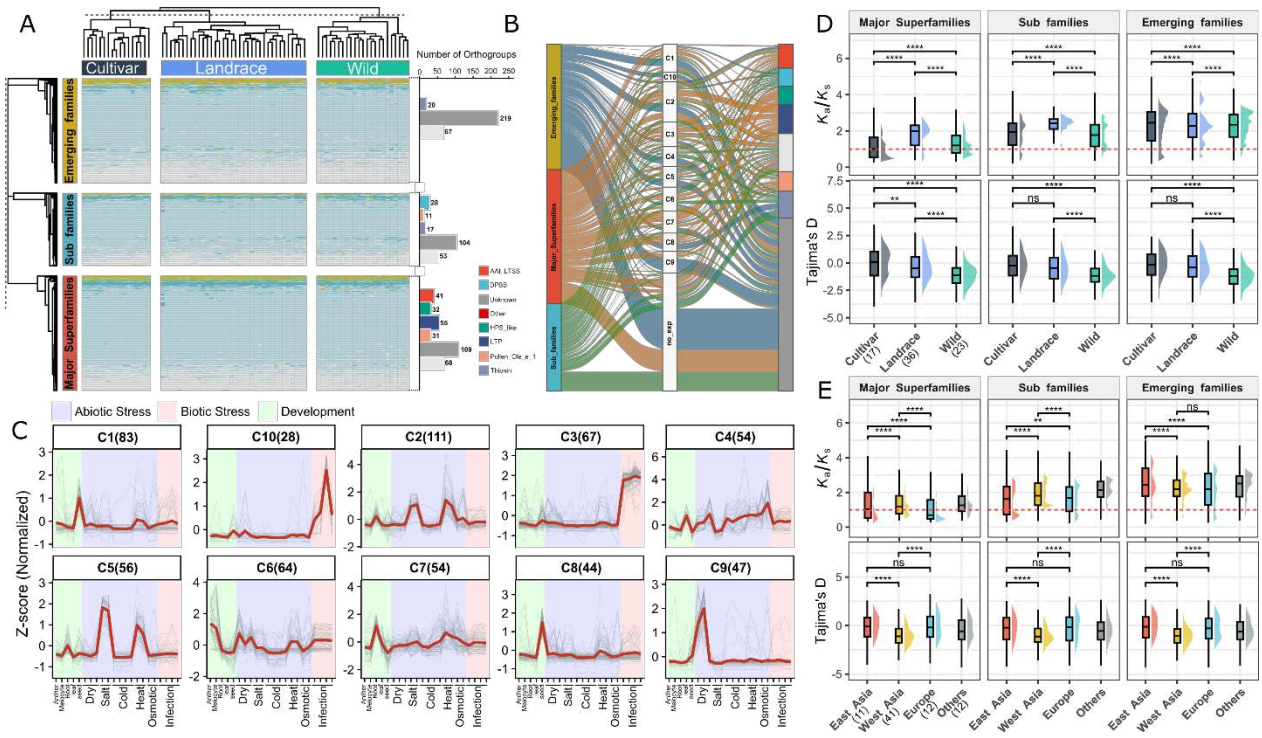
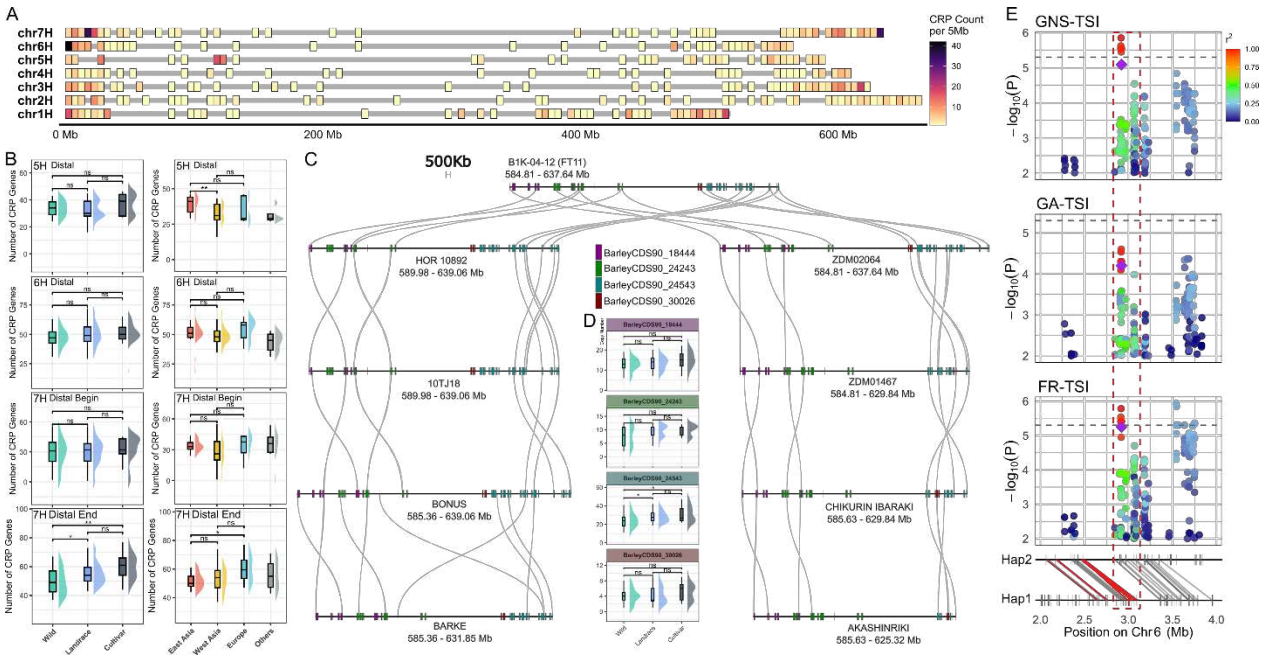


Fig. 4 Pangenome, transcriptome, and population genetic analyses of barley. (A) Presence–absence variation (PAV) heatmap of HC CRPs across 76 barley accessions in the pangenome. (B) Sankey diagram illustrating the relationships among structural classes (MS, SF, EF), expression clusters, and CDD-based functional annotations. (C) Expression clustering of HC CRPs, showing ten distinct clusters corresponding to developmental stages and treatment conditions. (D) Comparative analysis of Ka/Ks ratios and Tajima's D values for MS, SF, and EF across cultivar, landrace, and wild barley groups. Statistical significance was assessed using Student's t-test. (E) Comparative analysis of Ka/Ks ratios and Tajima's D values across geographic populations (East Asia, West Asia, and Europe; "Other" includes regions such as the United States). P-values from pairwise comparisons were adjusted across all 9 tests using FDR correction. FDR-adjusted $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$.



827

828 **Fig. 5 Candidate breeding loci associated with HC CRP clusters on barley chromosomes 6H**
829 **and 7H.** (A) Genomic distribution of HC CRP clusters in the barley cv. Morex reference genome.
830 (B) Abundance of HC CRPs within the four largest clusters across different domestication groups
831 and geographic regions. Statistical significance was assessed using Student's t-test. (C) Collinearity
832 of HC CRP clusters located in the distal region of chromosome 7H across representative genomes.
833 (D) CNVs of four orthogroups (OGs) in the distal region of chromosome 7H. Statistical
834 significance was assessed using Student's t-test. (E) Shared GWAS peaks for three traits located
835 in the proximal region of chromosome 6H, together with haplotype collinearity between
836 representative Hap1 and Hap2.

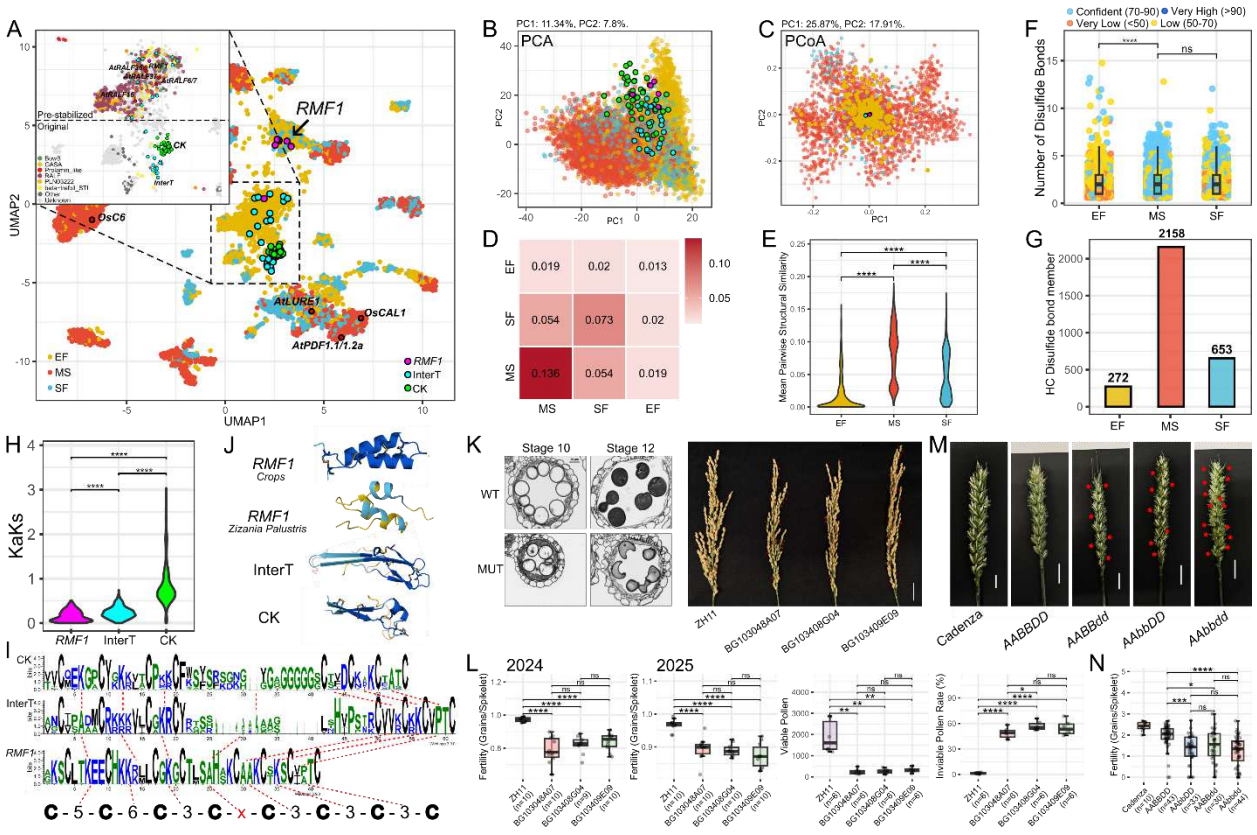
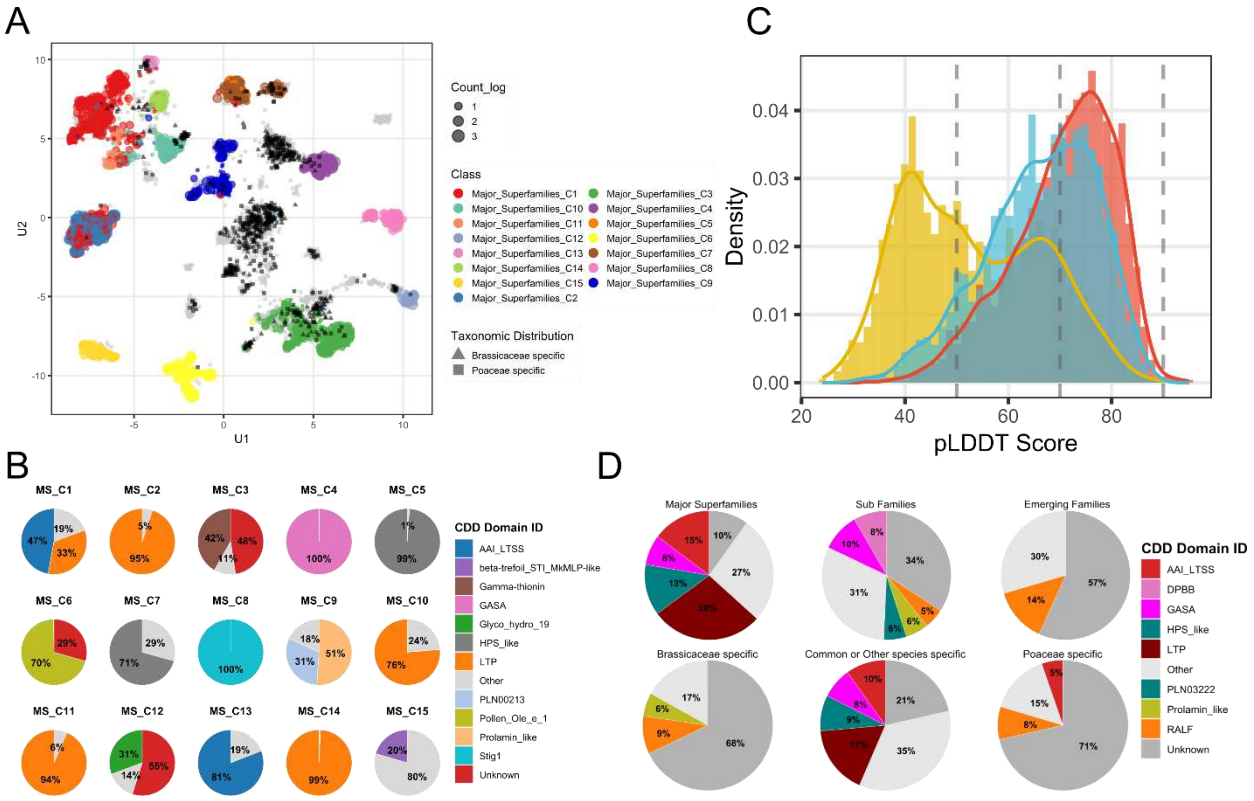
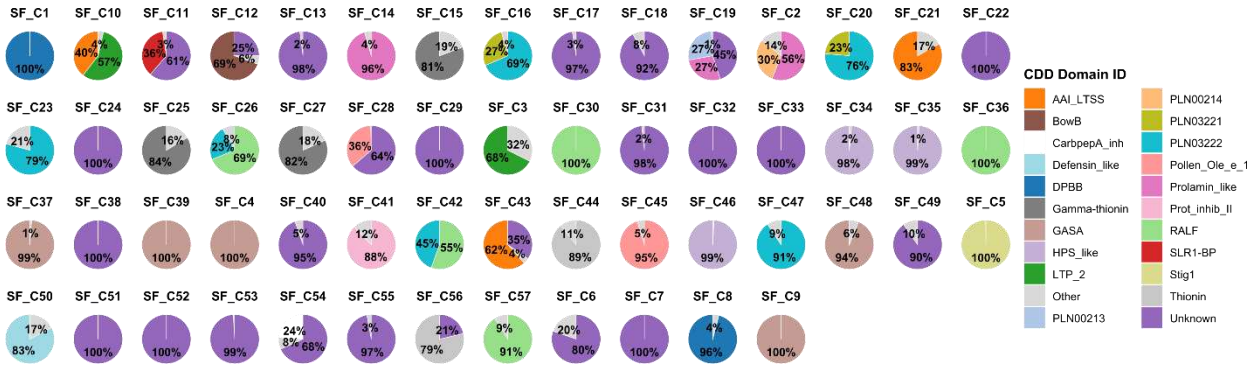


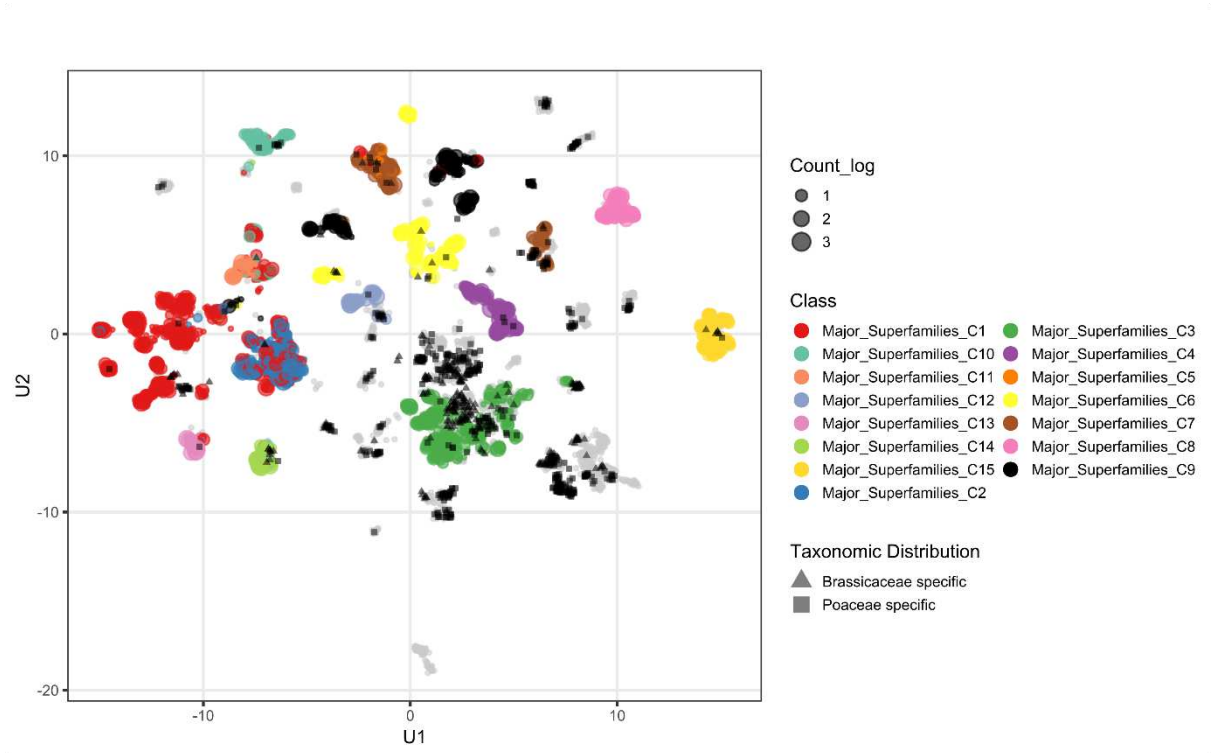
Fig. 6 The evolutionary reservoir hypothesis and functional characterization of the RMF1 lineage. (A) Structural landscape of the CRP superfamily and the evolutionary rsseservoir model. UMAP visualization across 868 species showing a central cluster of EFs as a “evolutionary reservoir”; a zoomed view of this "evolutionary reservoir" is also displayed. (B) PCA plot. Each point represents a protein structure. Colors indicate protein family classification. (C) PCoA plot. Each point represents a protein structure. Colors indicate protein family classification. (D) Heatmap of inter-family divergence and within-family compactness. The larger the number, the farther the average distance. (E) Pairwise structural similarity across MSs, SFs and EFs. (F) Count of disulfide bonds in EFs, MSs and SFs, statistical significance was assessed using student's t-test. (G) Count of member with HC disulfide bond in EFs, MSs and SFs. (H) Selective constraints on transitioning CRP lineages. Comparison of Ka/Ks ratios among CK, InterT, and RMF1, statistical significance was assessed using student's t-test. (I) Conserved cysteine frameworks. Sequence alignment of cysteine-rich domains from RMF1, InterT, and CK showing a conserved motif. (J) Structural metamorphosis predicted by AlphaFold3. Colored by default AlphaFold confidence pLDDT color code. Bule indicates high confidence. (K) Phenotypic effects of RMF1 loss-of-function. Representative images of panicles and dissected anthers from CRISPR/Cas9 *rmf1* knockout lines and wild-type (ZH11). (L) Quantitative analysis of reproductive traits. Seed setting rate, pollen number per anther, and pollen viability shown as mean ± SEM from independent biological replicates across two growing seasons (2024–2025), statistical significance was assessed using student's t-test. (M) Spike phenotype of wheat TILLING mutant in F2 population. (N) Quantitative analysis of fertility, statistical significance was assessed using student's t-test.



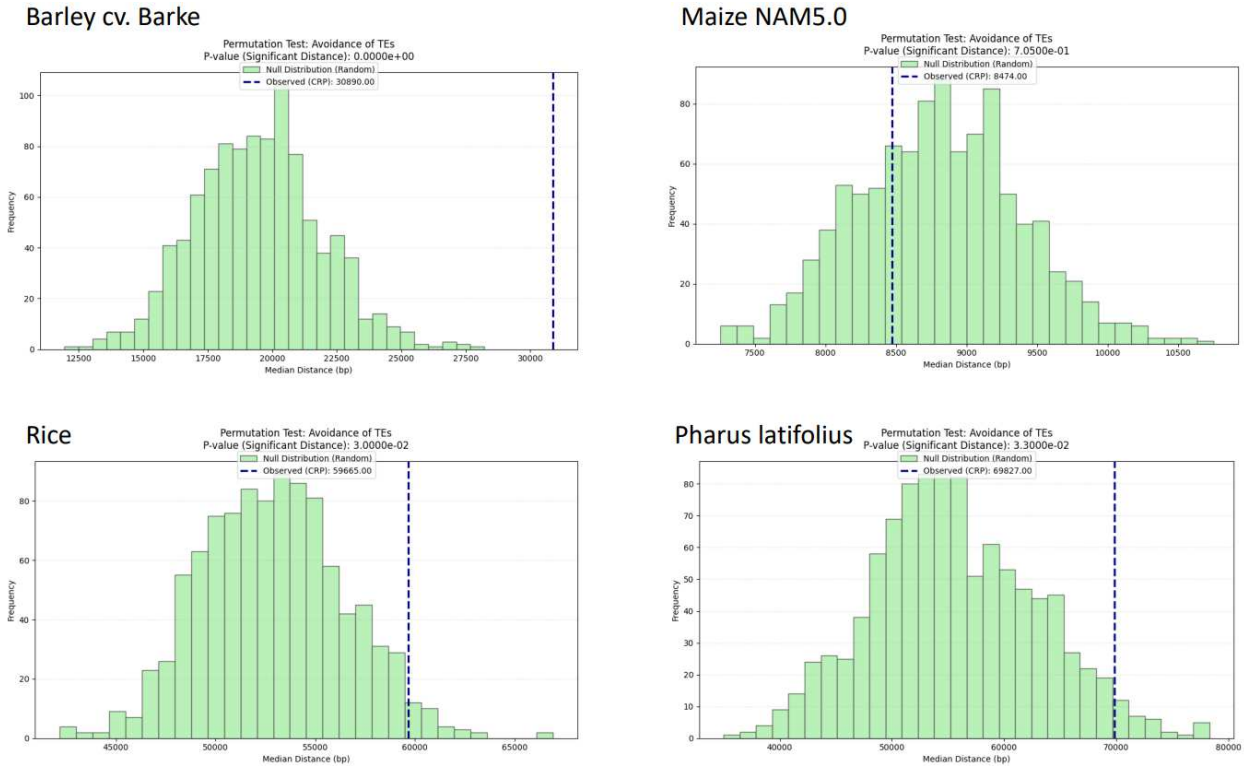
Supplementary Fig. S1. UMAP projection and taxonomic distribution based on structural similarity. (A) UMAP visualization based on TM-scores from structural alignments, annotated by MSs membership and taxonomic origin. (B) Distribution of conserved domain (CDD) percentages across different MSs clusters. (C) individual distribution of pLDDT score for EFs, MSs and SFs. (D) Functional annotation of HC CRPs based on conserved domain database (CDD) assignments.



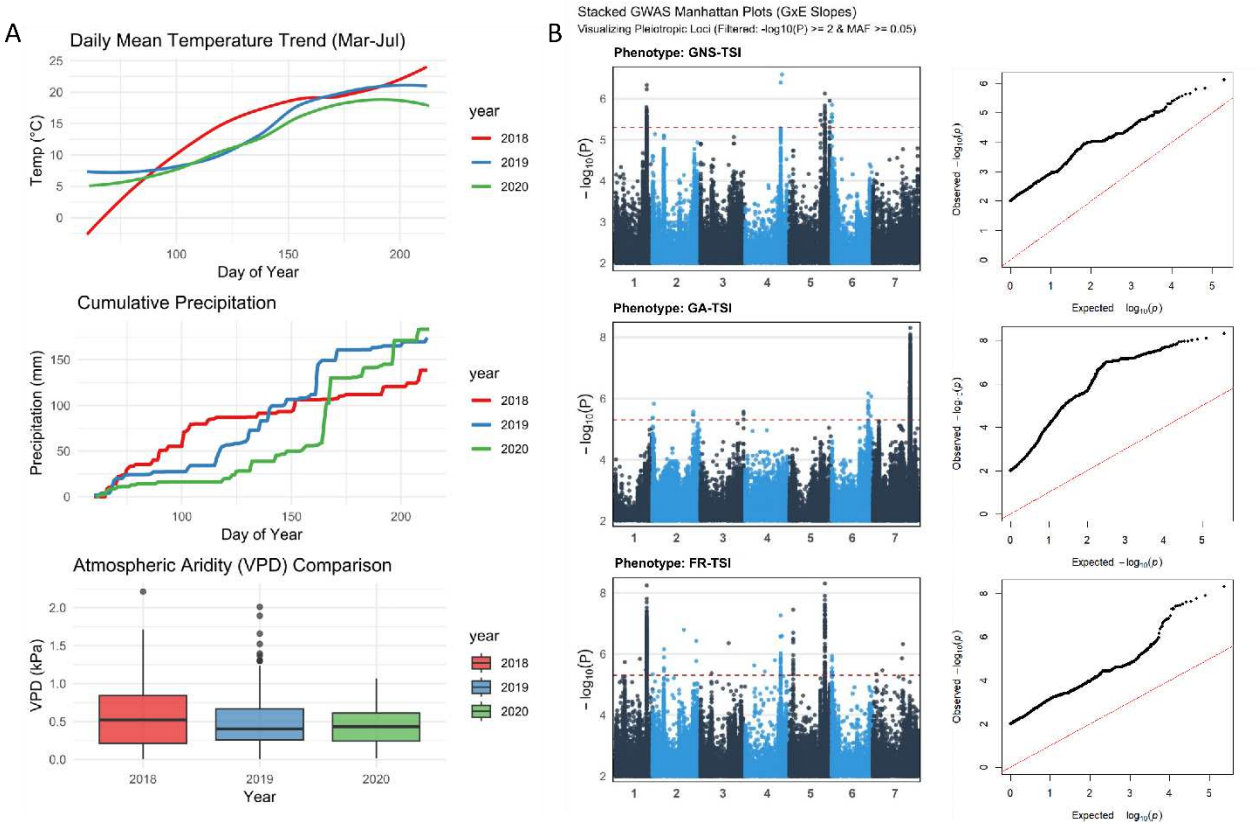
Supplementary Fig. S2. Conserved domain compositions within structural families (SFs). Percentage of CDD domains across different SFs clusters, highlighting domain-level conservation within structural groupings.



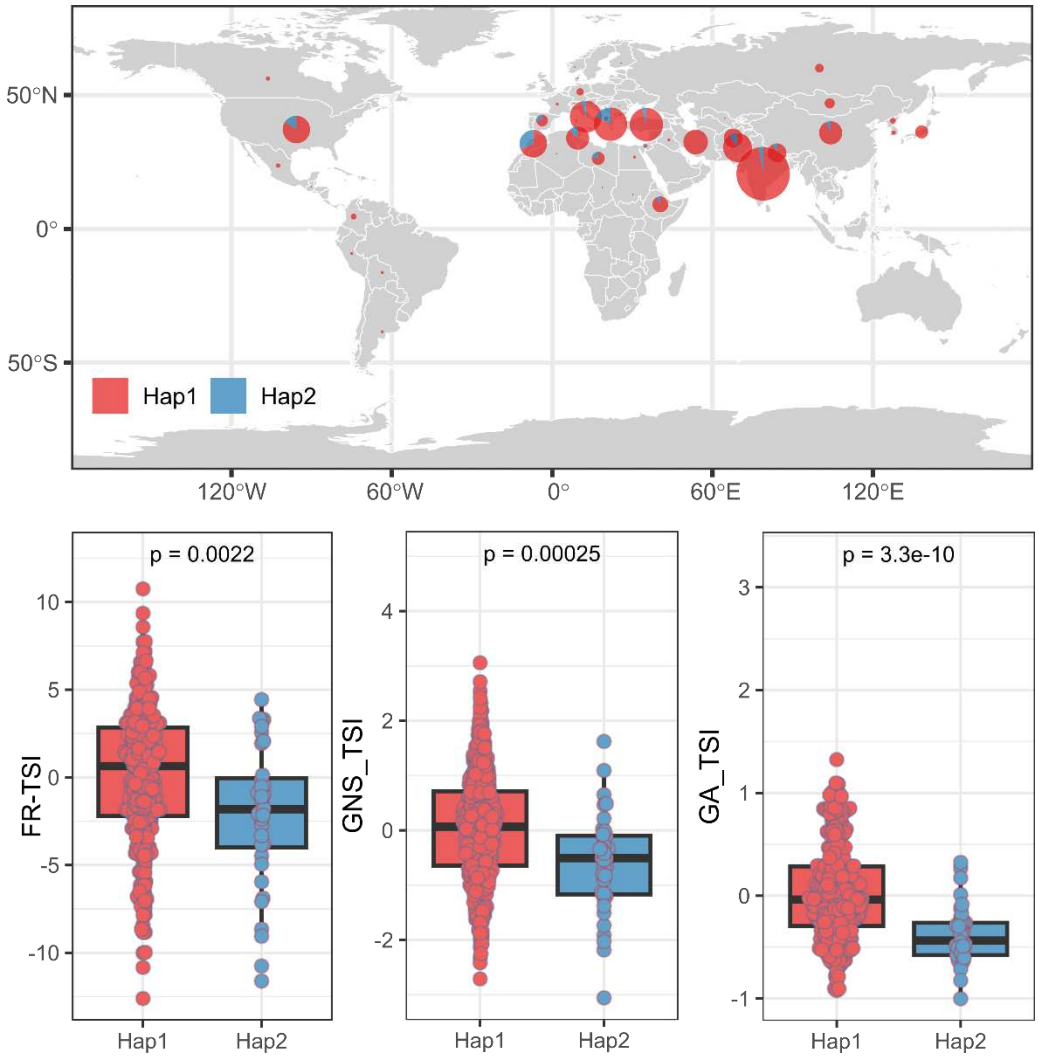
Supplementary Fig. S3. UMAP visualization based on sequence homology. Projection generated from normalized BLASTP bit-scores, with annotations for MSs and taxonomic distribution for comparison with structural clustering.



Supplementary Fig. S4. Genomic proximity of *CRP* genes to transposable elements (TEs). Comparison of distances to the nearest TE between *CRP* genes and a random background (1,000 permutations) across four species.

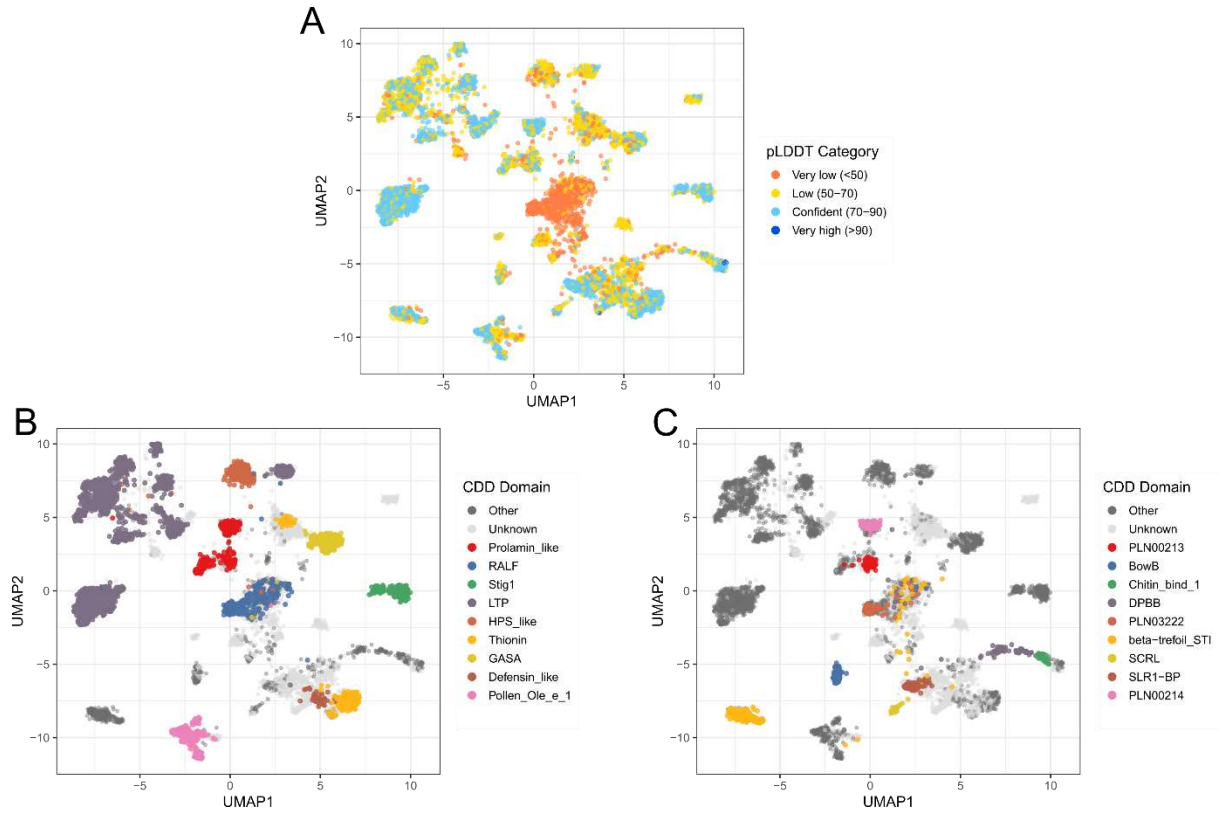


Supplementary Fig. S5. GWAS panel for environmental and trait associations. (A) Environmental variables including temperature, precipitation, and vapor pressure deficit (VPD) recorded from 2018 to 2020. (B) Manhattan plots and QQ plots illustrating the GWAS results for different TSI.



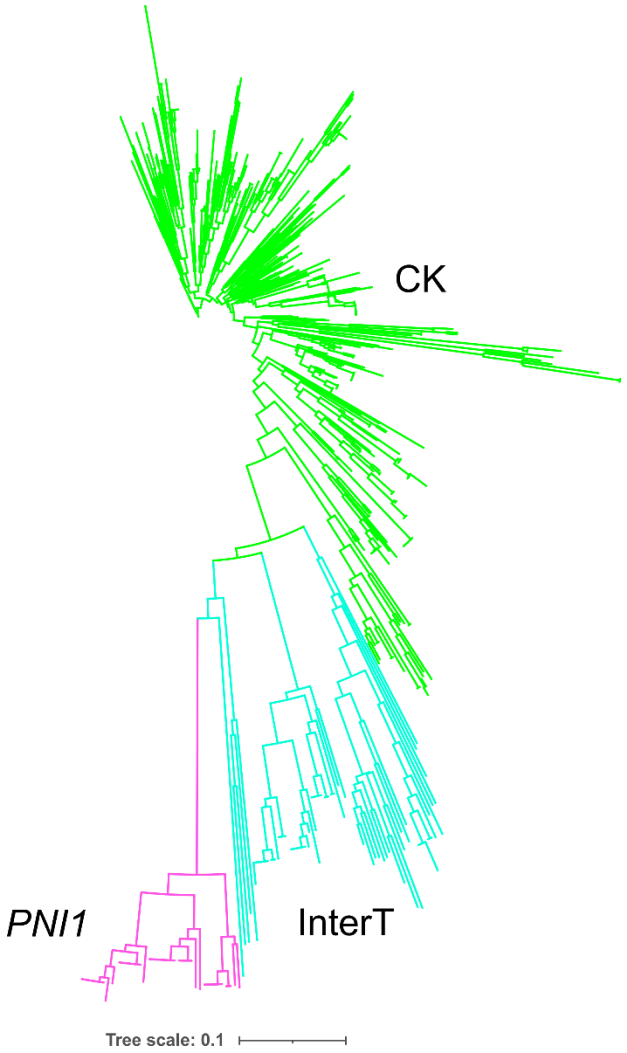
882

883 **Supplementary Fig. S6. Haplotype analysis of the candidate locus on chromosome 6H.** The panel
884 illustrates the geographic distribution of major haplotypes and the significant differences in different TSI
885 between Hap1 and Hap2.

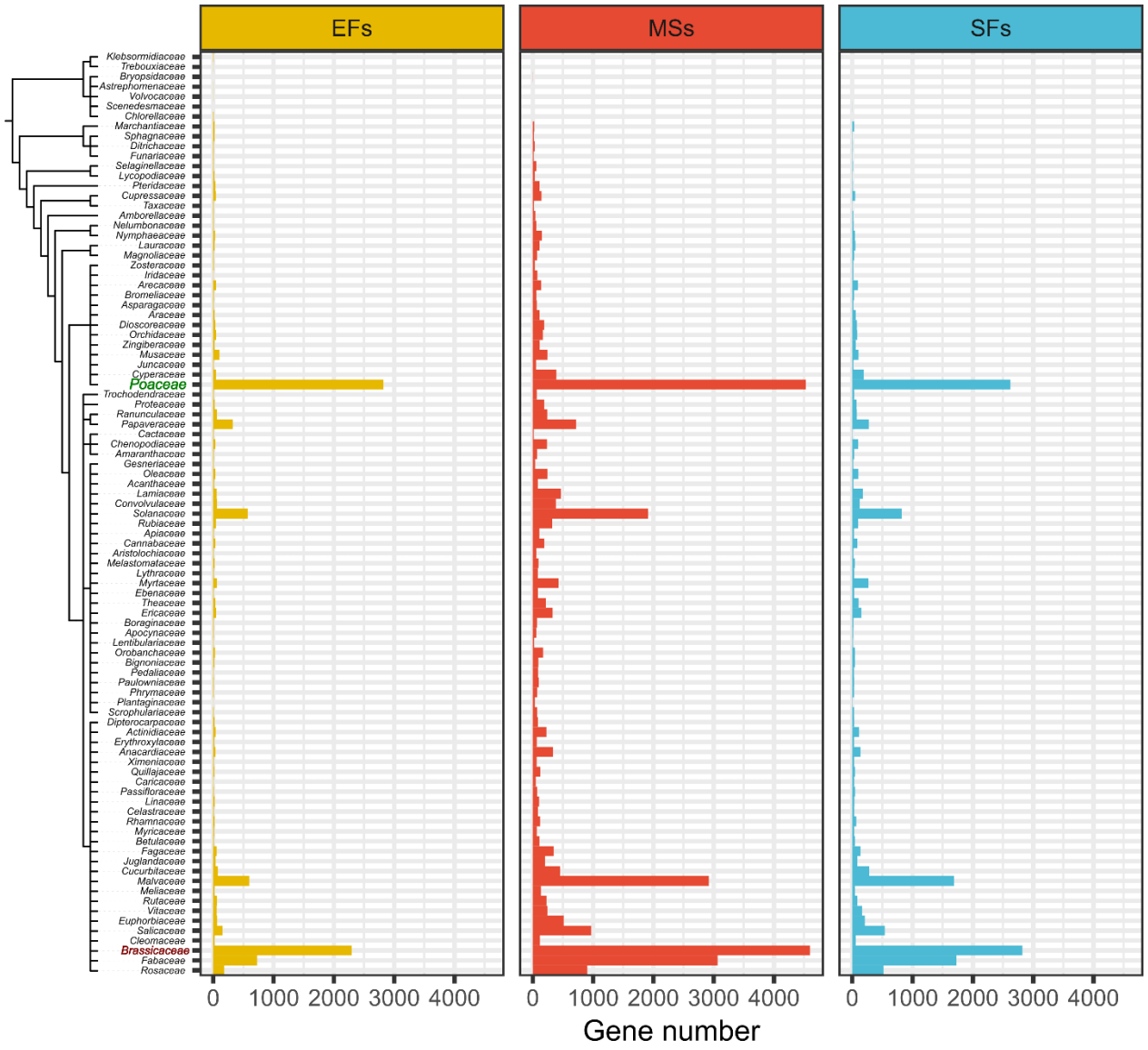


886

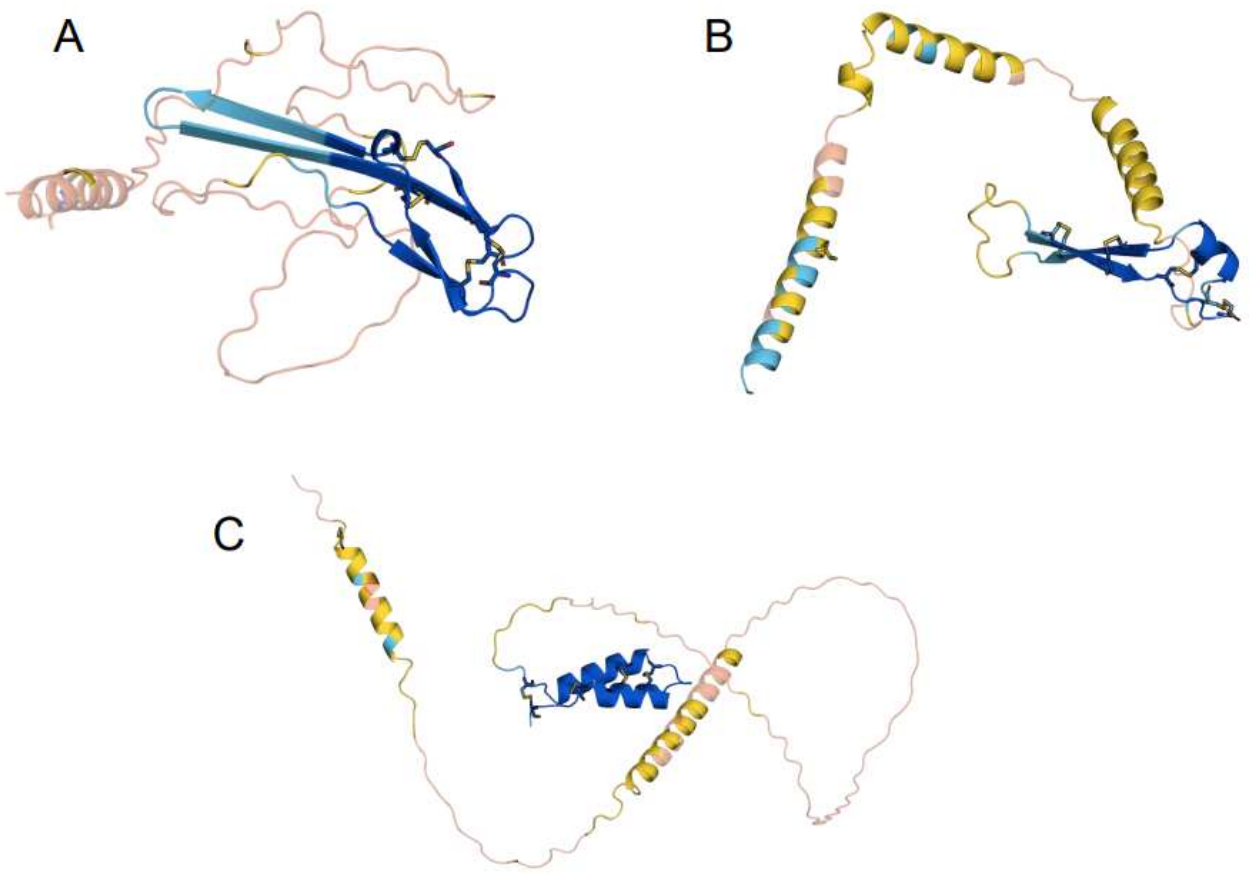
887 **Supplementary Fig. S7. Additional UMAP embeddings of the CRP structural landscape.** (A) UMAP
888 projection colored by pLDDT scores to contrast structural rigidity and intrinsic disorder. (B) UMAP
889 projection colored by Conserved Domain Database (CDD) annotations. (C) UMAP projection colored by
890 other CDD annotations.



Supplementary Fig. S8. Phylogenetic relationships of the CK, InterT, and RMF1 lineages. Phylogenetic reconstruction inferred via the neighbor-joining method based on core cysteine skeleton sequences. Taxa are color-coded by evolutionary strata: green highlights the basal CK elements, blue indicates the intermediate InterT elements, and pink denotes the structurally solidified RMF1 clade.



Supplementary Fig. S9. Evolutionary expansion profiles of EFs, MSs, and SFs across the plant kingdom. The cladogram (left) represents the organismal phylogenetic relationships. The horizontal bar plots (right) illustrate the total gene counts for EFs (yellow), MSs (red), and SFs (blue) within each corresponding plant family. The two major families undergoing catastrophic lineage-specific expansions are highlighted in colored text (*Poaceae* in green; *Brassicaceae* in red).



Supplementary Fig. S10. Structural characterization of CK, InterT, and RMF1 modeled by AlphaFold3. Three-dimensional conformational folds are shown for (A) CK, (B) InterT, and (C) RMF1. The structures are color-coded by the standard AlphaFold3 confidence metrics, ranging from high confidence (dark blue, pLDDT >= 90) to low confidence/intrinsic disorder (orange, pLDDT < 50).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementaryData.xlsx